Patient-oriented outcome assessment in rheumatic diseases

PATIENT-ORIENTED OUTCOME ASSESSMENT IN RHEUMATIC DISEASES

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Bakker, Carla Heleen

Patient-oriented outcome assessment in rheumatic diseases
/ Carla Heleen Bakker. - Maastricht : Universitaire Pers
Maastricht. - III.
Thesis Rijksuniversiteit Limburg Maastricht. - With ref.
ISBN 90-5278-188-5
Subject headings: rheumatology / utility assessment

This thesis was financially supported by grants from Ciba-Geigy, Glaxo, Kabi Pharmacia, Pfizer and "Het Nationaal Reumafonds".

PATIENT-ORIENTED OUTCOME ASSESSMENT

IN RHEUMATIC DISEASES

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. Mr. M.J. Cohen,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen
op vrijdag 16 juni 1995 om 12.00 uur

door

Carla Heleen Bakker

UPM
UNIVERSITAIRE PERS MAASTRICHT

Promotores:

Prof.Dr. JMJP van der Linden
Prof.Dr. EKA van Doorslaer (Universitaire Instellingen Antwerpen, België)

Beoordelingscommissie:

Prof.Dr. F Sturmans (voorzitter)
Prof.Dr. P Knipschild
Prof.Dr. RM Leidl
Prof.Dr. JJ Rasker (Medisch Spectrum Twente)
Prof.Dr. P Tugwell (University of Ottawa, Canada)

Hé-Ho, Hé-Ho,
je krijgt het niet kado ...
(Sneeuwwitje en de zeven dwergen)

Aan Berno, Floris en Ruben
aan mijn ouders

CONTENTS

1

INTRODUCTION: GUIDE TO THESIS

INTRODUCTION: GUIDE TO THESIS

Interest in patient-oriented outcome assessment in rheumatic diseases has increased in recent years. Measuring outcome or health status is important for assessing the impact of chronic diseases on the patient.[1,2] Outcome is the net effect, end result or endpoint and may be divided into objective measures of disease activity assessed by the physician and the subjective outcomes based on the patient's report.[3]

To measure outcome comprehensively it should include all those components of health status important to patient and physician that are relevant to the intervention assessed.[3] Health status can be described in 5 dimensions or domains, frequently abbreviated as 5 D's (modified from White)[4-6]: Death, Disability, Discomfort, Drug (or therapeutic toxicity), Dollar costs. Tugwell suggested 8 D's (modified from White)[3]: Death, Disease activity, Distress, Disadvantages (drug or therapeutic toxicity), Disability (and Dysfunction), Disharmony, Dissatisfaction. Some dimensions might be more easily assessable than others. For example, Death is easy to assess, (however, an infrequent endpoint in chronic rheumatic diseases), whereas Disability is more difficult to assess and requires a variety of measures.

The term outcome is frequently used interchangeable with health related quality of life or health status. The scope of health status can be either specifically oriented, focussing on only one dimension, or broadly oriented, focussing on several dimensions. Further, different diseases may deduct different aspects of health status, whereas individual patients may have different priorities regarding their health status. Physician assessed and patient reported measures can be classified as either specific or generic in each of 3 areas of focus: health status, disease and patient (Figure 1). A measure is specific in the health status area when it measures only one dimension, and generic when it measures several dimensions. A measure is specific in the disease area when it is applicable to only 1 disease (e.g., ankylosing spondylitis). It becomes less focussed when the measure can be applied in a group of diseases (e.g., all arthritides, all cancers) and completely generic when applicable to all diseases. Similarly, measures that are specific in the patient area refer to single patients. Less specifically focussed measures refer to subgroups (e.g., the elderly) and generic measures refer to all possible patients. Obviously, outcome measures are multidimensional and can be generic in one area and focussed in another.

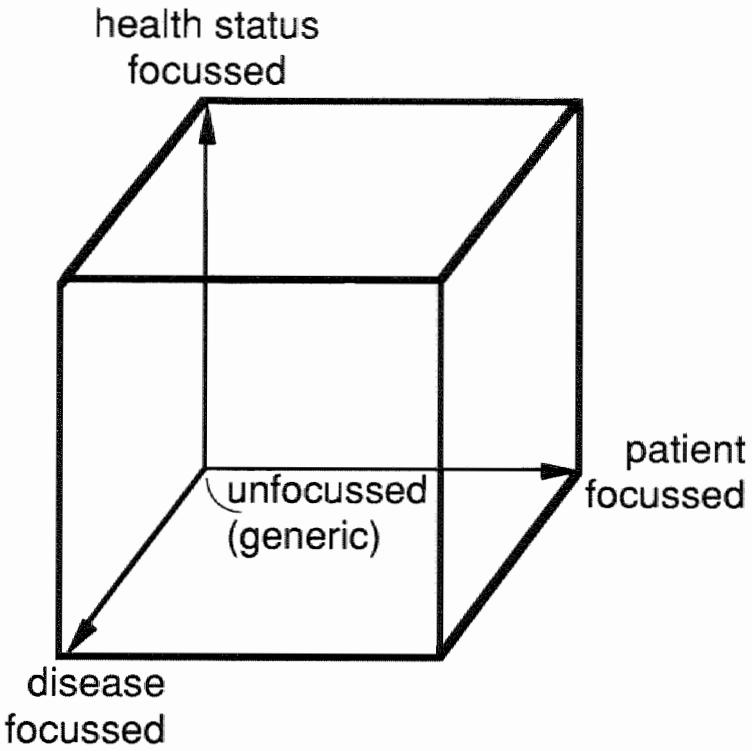


Figure 1 Taxonomy of physician assessed and patient reported measures regarding the three areas or directions of focus: health status, disease, and patient

Measures used today in rheumatic diseases do not often cover the whole spectrum of health status. This thesis focusses on patient-oriented measures in 2 rheumatic diseases: ankylosing spondylitis and fibromyalgia. Both the use of such instruments and the development and testing of patient-preferences are studied. The relationships between the costs and effects of an intervention are also examined. **Chapter 2** studies the current use of instruments - both patient reported and physician ordered or physician assessed - in ankylosing spondylitis.
Chapter 3 provides an introduction to utility measurement in rheumatology. Utility measures are generic measures (focussing on several D's) assessing priorities of individual patients. Utilities assess the value or preference a patient attaches to his (her) overall health status. In these measures patients summarize the risks and benefits of an intervention into one overall single value that allows comparison of outcomes across patients between various health care interventions and different health states or diseases. As yet, only one published randomized controlled drug trial in the field of rheumatology has used utility measures. This trial concerned the evaluation of auranofin in rheumatoid arthritis patients.[7] We argued that this unique priority measure needed further application in clinical trials in rheumatic diseases. We elicited patient priority values from ankylosing spondylitis and fibromyalgia patients in three randomized controlled trials: 1) a study on the comparison of two NASIDs in ankylosing spondylitis

patients; 2) a study on the overall therapeutic effect of low-impact fitness training and biofeedback training in fibromyalgia patients; 3) a study on the effectiveness of supervised group physical therapy compared to unsupervised exercises at home in ankylosing spondylitis patients.

Chapter 4 describes the feasibility of utility measurement by rating scale method and standard gamble method in ankylosing spondylitis and fibromyalgia patients, whereas **chapter 5 and 6** describe the association of utility measures with other outcome measures in these patients respectively. **Chapter 7** provides a detailed description of methodological issues of utility measurement.

Another way to assess patient preferences, besides utilities, is by means of the Problem Elicitation Technique (PET) questionnaire.[1] This newly developed instrument was specifically designed to assess outcome (disability, discomfort and drug toxicity) focussed to the individual patient. It deals only with activities that are directly limited by the disease and judged important by the patient.[8] This may improve the responsiveness to clinically relevant improvement while reducing the sample size needed in clinical trials. We applied the PET questionnaire to both ankylosing spondylitis and fibromyalgia patients. The results in terms of construct validity and sensitivity to change are described in **chapter 8**.

Another aspect of outcome - the D of Dollar costs was addressed in a cost-effectiveness study of supervised group physical therapy compared to unsupervised exercises at home in patients with ankylosing spondylitis (**chapter 9**).

Finally, **chapter 10** gives a general discussion. The thesis ends with summaries in English and Dutch.

REFERENCES

1 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.

2 Guyatt GH, Feeny DH, Patrick DL: Measuring health-related quality of life. Annals Internal Med 1993; 118:622-9.

3 Tugwell P, Bombardier C: A methodological framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982; 9:758-62.

4 Fries JF: Toward an understanding of patient outcome measurement. Arthritis Rheum 1983;26:697-704.

5 Fries JF, Spitz P, Kraines RG, Holman HR: Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.

6 White KL: Improved medical care: statistics and the health services system. Public Health Rep 1967; 82:847-54.

7 Bombardier C, Ware J, Russell J, Larson M, Chalmers A, Read, L: Auranofin therapy and quality of life in patients with rheumatoid arthritis. Am J Med 1986; 81:565-78.

8 Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B: The MACTAR patient preference disability questionnaire: an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987; 14:446-51.

2

MEASURES TO ASSESS ANKYLOSING SPONDYLITIS: TAXONOMY, REVIEW AND RECOMMENDATIONS

Carla Bakker, Maarten Boers, and Sjef van der Linden

Department of Internal Medicine, Division of Rheumatology, University of Limburg, Maastricht, The Netherlands

Reprinted by permission of the Journal of Rheumatology 1993;20:1724-30

MEASURES TO ASSESS ANKYLOSING SPONDYLITIS: TAXONOMY, REVIEW AND RECOMMENDATIONS

ABSTRACT

Objective. To critically review the current use and scope of measures to assess patients with ankylosing spondylitis (AS).

Methods. Studies in English reported between January, 1986 and August, 1991 were identified both through computer searches of *Index Medicus* and manual searches of bibliographies. Only studies where assessment of AS was a main topic were included. Information was extracted to classify measures as 1) physician assessed, 2) patient reported or 3) other assessments.

Results. Physician assessed measures prevailed in 34 (79%) of the 43 studies included. Patient reported measures were mentioned in 29 (67%). Most physician assessed measures (67%) focussed on mobility, most patient reported measures (65%) focussed on discomfort. Single item global assessment by physician or patient, the most generic measure, was reported in 7 (16%) and in 17 (40%) studies, respectively. One study reported a measure which specifically addressed the patient's priorities regarding treatment risks. Other measures were reported in 22 (51%) studies, i.e., laboratory tests in all 22, and additionally radiographs in 2, and various measures in 6 studies. Side effects (by reports or otherwise) were noted in 26 (60%) studies.

Conclusion. Current assessment in AS incompletely encompasses the spectrum of relevant health status outcomes. Specifically, more attention should be paid to the patient's point of view.

INTRODUCTION

We review the measures available to assess changes over time in patients with ankylosing spondylitis (AS).[1] These changes are due to the course of the disease and the effects of therapeutic interventions.

Attempts have been made to categorize measures as process or outcome.[2,3] *Outcome* is the net effect or end result, whereas *process* indicates what happens along the way.[2] For example, death or disability is an outcome measure and laboratory measures such as sedimentation rate qualify as process measures. Outcome measures are usually assessed by physicians or patients. They are defined by five dimensions, i.e., death, disability, discomfort, drug (or therapeutic) toxicity, and cost.[2,3] These dimensions are not mutually exclusive.

In recent years a variety of instruments have been developed to measure outcome as perceived by patients. Most of these have been applied to patients with rheumatoid arthritis. Examples include the Sickness Impact Profile (SIP), the Health Assessment Questionnaire (HAQ) and the Arthritis Impact Measurement Scale (AIMS).[3-5]

Frequently, the term outcome is used interchangeably with the term health status. The scope of *health status* can either be specifically oriented, focussing on only one dimension, or broadly oriented (generic), focussing on several dimensions. Further, different *diseases* may deduct different aspects of health status, whereas individual *patients* may have different priorities regarding their health status. Clearly, this may have implications for the instruments used to assess health status in patients with AS. Our purpose is to critically review the current use and scope of measures to assess patients with AS.

MATERIALS AND METHODS

We searched by CD-ROM through the Medline database using different combinations of textwords (i.e., ankylosing spondylitis, spondyloarthropathy, sacroiliitis, trial, therapy) and exploded keywords (i.e., spondylitis, clinical trials, outcome and process assessment, health surveys, questionnaires). We also searched bibliographies of relevant articles. The search comprised articles and letters reported in English between January, 1986 and August, 1991. It included all methodological articles (e.g., describing the development of a new instrument), surveys and clinical trials having assessment in AS as a main topic. We did not apply specific methodological guidelines to the articles obtained.

We categorized measures as physician assessed, patient reported or "other", including laboratory tests and radiographs. The measures intend to assess health status at a given point in time, although some are applied to monitor disease activity or safety rather than health status. We applied a taxonomy, which is a refinement of the taxonomy of Guyatt et al.[6] In this taxonomy a measure is classified as either specific or generic in each of 3 areas of focus: *health status, disease* and *patient* (Figure 1). A measure is specific in the *health status* area when it measures only one dimension, and generic when it measures several dimensions. A measure is specific in the *disease* area when it is applicable to only 1 disease (e.g., AS). It becomes less focussed when it is applied in a group of diseases (e.g., all arthritides, all cancers) and completely generic when applicable to all diseases. Similarly, measures that are specific in the

patient area refer to single patients. Less specifically focussed measures refer to subgroups (e.g., the elderly) and generic measures refer to all possible patients.

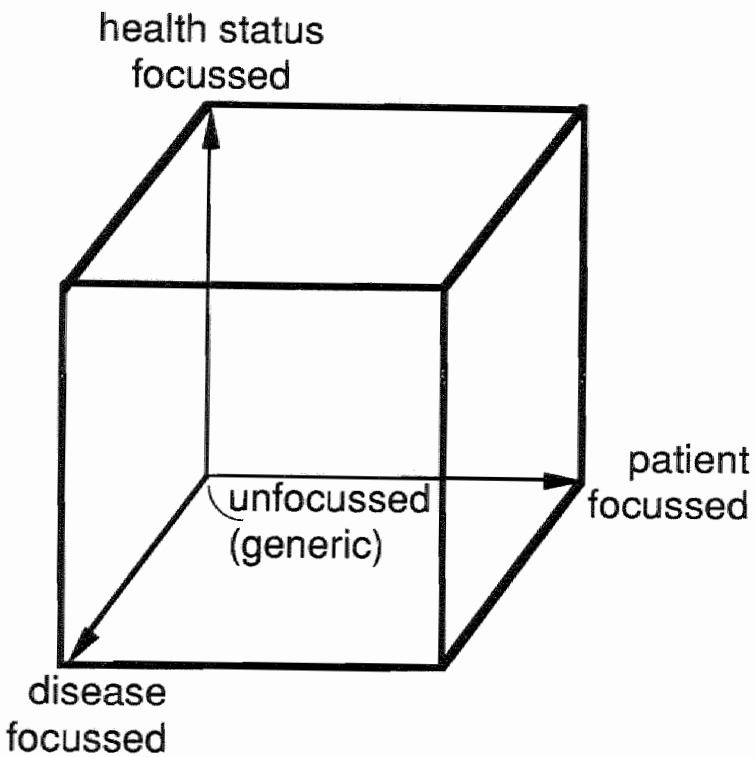


Figure 1 Taxonomy of physician assessed and patient reported measures regarding the three areas or directions of focus: health status, disease, and patient

Obviously, a measure can be generic in one area and focussed in another. For example, utility measurements applied to individual patients are usually focussed on the patient area as they measure the patient's individual priorities, but generic in the health status area.
We considered compliance under "other tests". Compliance by itself is not a direct measure of effect, but an effect modifier. Compliance will not be categorized in our review but mentioned separately.

RESULTS

Of a total of 260 titles, 43 articles and 4 letters [7-10] were included in the review (Table 1). Four articles reported new instruments to assess AS.[11-14] Twenty-eight articles (14 randomized controlled trials) and all 4 letters evaluated treatment, either with drugs [7,8,10,15-38], physical therapy [39-42] or acupuncture.[9] The other 11 reports did not deal with intervention, but were case-control [43,44], cohort [45], and cross sectional studies [46-53] (Table 1).

Table 1. Review of 47 reports on assessment of patients with ankylosing spondylitis

A.	Development of new instruments (n=4):	
	1. Functional and Articular Index	
	2. Enthesis Index	
	3. Health Assessment Questionnaire for Spondyloarthropathies	
	4. A simple method for measuring lateral flexion of the dorsolumbar spine	
B.	Evaluation of an intervention (n=32):	
	27	Drug trials - 17 NSAID - 10 second line drugs
	4	Physical therapy trials
	1	Acupuncture trial
C.	Case-control study (n=2) and cohort study (n=1)	
D.	Cross-sectional studies (n=8)	

We found 3 instances of overlapping publication, for example several aspects of the same study or an expanding study. Six articles were all based on the same sample of AS patients who had returned a mailed questionnaire.[43,44,46-48,52] A comparative trial of nonsteroidal antiinflammatory drugs (NSAID) reported on disease severity and pulmonary function in 2 separate articles.[23,24] One trial was reported both in a letter and an article.[10,17] Each instance of overlapping publication was counted as one study in our review.
Of the 43 articles 13 were reported in 1986, 4 in 1987, 7 in 1988, 6 in 1989, 13 in 1990 and 0 in 1991. Thirty-four articles were from Europe, 7 from the USA, 1 from South America, and 1 from Australia. Three letters came from Europe and 1 from South America.

Review of measures reported in 36 articles

All but 2 studies reported *physician assessed* measures. These measures included spinal mobility (assessed with Schober test, fingertip to floor distance, chest expansion, occiput to wall distance), joint and enthesis pain on examination (assessed with articular index, painful or swollen joint count and number of tender enthesis), and single item global assessment by physician (Table 2). This is the most generic measure in the health status, disease, and patient area. The terms 'global', 'general' and 'overall' were used interchangeably, and may have

comprised different evaluations of therapy, i.e., efficacy, disease activity, health status, and toxicity.
Drug toxicity was frequently reported, usually by simple tabulation of physician assessed side effects. The physical therapy trials did not report adverse effects of treatment. The single letter on acupuncture also did not mention side effects.
All physician assessed measures were classified as patient generic, and except for the global assessment, as health status and disease focussed. In particular, the measures focussed on the disability (mobility), discomfort (pain), and drug toxicity health status dimensions. The target disorder was always AS and applied to all patients with this disease.

Table 2. Physician assessed measures used in 34 studies *

	Number of times used	Focus in disease area **
DISABILITY		
(Spinal) Mobility (29 studies)		
(Modified) Schober	24	+
Fingertip to floor distance	21	+
Chest expansion	22	+
Occiput to wall distance	12	+
Cervical rotation	3	+
Hip mobility	3	+
Other ***	12	+
	97	
DISCOMFORT		
Pain / Tenderness (18 studies)		
Spinal Tenderness	8	±
Peripheral Painful joint count	12	±
Swollen joint count	5	±
Articular index	2	+
Number of tender entheses	3	+
	30	
DRUG (measured in 26 studies)		
Adverse reactions	26	-
Evaluation of tolerance by physician	1	-
Global safety by physician	2	-
	29	
GLOBAL		
General or overall assessment (7 studies)		
Assessment by physician	10	-

* Overlapping articles counted as one study. All reported measures were classified as health status focussed except the global assessment. All measures were classified as patient generic.

** + = focussed in disease area
− = generic in disease area

*** Other: range of spinal motion[9,18,28,34], cervical flexion[30], chin to chest distance[22], Wall tragus distance[7], Otto's test[17], height[41], modified foot measure[39], spirometric examination[23,24], temporomandibular joint examination.[42]

Thirty-four studies used at least one *patient reported* measure, most frequently pain or stiffness (Table 3). Many studies used multiple measures for the same aspect (e.g., pain). With one exception [49], all patient reported measures were patient generic but health status and disease focussed. In particular, they focussed on the discomfort dimension and on arthritis as a group of diseases. Only a few studies included patient reported measures to cover the disability and the cost dimensions. A total of 17 studies reported global assessment by patient (Table 3).
In one study, a standard gamble method assessed the patient's individual willingness to accept treatment risks.[49] Because it assessed patient priorities, this measure is patient focussed. At the same time the standard gamble measure is disease generic.

Table 3. Patient reported measures used in 34 studies *

	Number of times used	Focus in disease area**
DISABILITY		
Functional status (9 studies)		
Functional index questionnaire	3	+
Steinbrocker functional class	2	+
Ankylosing Spondylitis Assessment Questionnaire	1	+
Toronto activities of daily living Questionnaire	1	+
Disability index of Health Assessment Questionnaire	1	+
50 foot walk time	1	+
	9	
DISCOMFORT		
Pain (28 studies)		
General - Severity only	14	−
- Analgesic consumption	5	−
Specific - Spine	16	±
- Peripheral joints	4	±
- Pain at night	10	−
- pain on movement	4	−
- pain at rest	2	−
Stiffness (28 studies)		
Morning stiffness	23	+
Severity of stiffness	8	+
Immobility stiffness	2	+
Sleep disturbance (due to pain or stiffness)	8	−
Time to onset of fatigue	5	−
	101	
DRUG or therapeutic toxicity (4 studies)		
General safety	1	−
General tolerance	1	−
Willingness to accept risk	1	−
Risk perception questionnaire	1	−
	4	
COMBINATION OF DISABILITY, DISCOMFORT and DRUG		
Global, general, overall assessment (17 studies)		
Assessment by patient	18	−
Satisfaction with physiotherapy program	1	−
Patient's desire to continue drug treatment	3	−
Health status measured in 2 studies		
modified AIMS	1	±
Nottingham Health Profile	1	−
	24	
DOLLAR costs (3 studies)		
Time off work	2	−
% people working	1	−
Sick leave	1	−
Payment of disability pension	1	−
Changes of work due to AS	1	+
	6	

* Overlapping articles counted as one study. All reported measures were classified as health status focussed except the global assessment and the 2 generic health status questionnaires. All measures were classified as patient generic except "willingness to accept risk" (drug toxicity).

** + = focussed in disease area
− = generic in disease area

Twenty-two studies reported *other tests* (Table 4). Laboratory measures to evaluate and monitor disease activity and side effects included blood counts (18 reports), acute phase proteins (17 reports), liver function (15 reports), renal function (14 reports), electrolytes or vitamins (5 reports), urinalysis (13 reports) and stool for occult blood (4 reports). In 2 studies radiological examinations were reported. Also, 4 studies included electrocardiograms, 5 trials reported ophthalmologic examination, 3 audiologic examination and 1 colonoscopy.
Compliance (assessed with tablet count, diary card, blood or urine samples) was assessed in 4 studies.

Table 4. Other measures used in 22 studies

	Number of times used
Laboratory measures (22 studies)	
Blood counts	18
Acute phase proteins	17
Liver function	15
Renal function	14
Other blood measures	5
Urinalysis	13
Stool for occult blood	4
	86
Radiographic examination (2 studies)	
Axial skeleton	2
Peripheral joints	1
Bone scan	2
	5
Other tests (6 studies)	
Electrocardiogram	4
Ophthalmologic examination	4
Audiologic examination	3
Colonoscopy	1
	12

Development of new measures

New *physician assessed* measures are the articular index[13], the enthesis index[14], a simple method for measuring lateral flexion of the dorsolumbar spine[12], and part of the index of disease activity.[23] The articular index measures discomfort. It is based on scoring of pain of 10 selected joints at movement or firm digital pressure.[13] Intra and interobserver reliability of the articular index are high ($r=0.83$ and $r=0.94$, respectively).[13] The index was able to detect exacerbation

and remission of symptoms during NSAID therapy.[13]
The enthesis index measures discomfort by a scoring system based on the patient's response to firm palpation over entheses easily accessible to examination.[14] The enthesis index is sensitive to change after one week of treatment with NSAID. This index correlates well with pain ($r=0.67$, $p<0.01$) and stiffness ($r=0.46$, $p<0.05$) scores.[14]
The new method for measuring lateral flexion of the dorsolumbar spine with a tape correlated well with assessments of lateral flexion with an inclinometer ($r=0.997$).[12] This new method of assessing disability has not yet been used in a published therapeutic trial.
An index of disease activity measuring discomfort and disability was used in one NSAID trial.[23] It is a composite measure calculated from 2 physician assessed (chest expansion and lumbar flexion index) and 2 patient reported measures (spinal pain and morning stiffness).[23]
All new physician assessed measures are patient generic, but health status and disease focussed.
New *patient reported* measures are the Functional Index[13], the Health Assessment Questionnaire for spondyloarthropathies (HAQ-S)[11], and the Ankylosing Spondylitis Assessment Questionnaire (ASAQ).[43,44,52] Additionally, 4 studies reported measurements with existing patient reported measures that had been modified slightly.[39,43-45,52,53]
The Functional Index measures disability and consists of 20 questions corresponding to activities of daily living. It takes about 2 min to complete. Intra and interobserver reliability are high ($r=0.86$ and $r=0.99$, respectively).[13] The index was able to detect exacerbation and remission of symptoms during NSAID therapy.[13]
The HAQ-S, assessing functional status, pain and stiffness, measures both disability and discomfort.[11] Five items were added to the original HAQ: carrying grocery bags, sitting for long periods of time, working at a flat topped table or desk, driving a car in reverse and using the rear view mirror. This modification raised the mean difficulty score from 0.38 (SD = 0.49) to 0.49 (SD = 0.51) on a 0 to 3 scale, indicating a slightly increased ability to capture functional limitations.[11] Neck rotation correlated most strongly with the HAQ-S score.[11] Sensitivity to change of the HAQ-S is as yet unknown, and it has not yet been reported in clinical trials.
Another study introduced the ASAQ.[43,44,52] This measure focusses on spinal mobility and pain. Its characteristics, including reliability, validity or sensitivity, have not yet been published. The same study also proposed a modified version of the Arthritis Impact Measurement Scale (AIMS).[43,44,52]
One study applied a slightly modified version of the Toronto Activities of Daily Living Questionnaire.[39] Another study assessed disability after modification of the disability index of the original HAQ.[45] Apart from minor modifications, 2 questions were added concerning turning the head left and right, and looking upwards.[45] Finally, 1 study used a questionnaire aimed at activities of daily living, problems when driving a car, and sexual problems.[53] No further details of this questionnaire were given.
All new or modified patient reported measures are patient generic, but focussed in both the disease and health status areas.

DISCUSSION

Clinical patient oriented research in AS is clearly a growing field. We document the wide variety of measures used today, but also their deficiencies. Currently, most measures in AS are focussed on the disease and health status area, and are generic in the patient area. It appeared

from our review that the disability dimension of health status in AS is highly represented in the physician assessed measures, and the discomfort dimension of health status in the patient reported assessments. Apart from single items of global assessments, general measures of health status were only occasionally used. In particular, self-assessed measurement of functional ability was frequently lacking. Drug toxicity reports missed informative detail. Finally, only 3 studies included an economic analysis. Therefore, some areas of the total spectrum of health status are not well covered by the measures used, whereas other areas are overrepresented.

We restricted our review to *what kind* of measures have been used in AS, and their classification. In other words, we were mostly concerned with aspects of credibility (face validity) and comprehensiveness (content validity). Other validity aspects have recently been discussed elsewhere.[54-58] Such aspects include selecting a measure appropriate for the purpose of the trial in terms of accuracy, reliability, and sensitivity to change. For example, assessment of spinal movement is an insensitive measure in patients with a fused spine. Other problems with validity include multiplicity and lack of standardization of existing measures. Multiplicity occurs when several measures of spinal mobility or joint pain counts are used within one trial. Lack of standardization is demonstrated by the remarkably large number of ways in which spinal mobility, general pain, joint pain, stiffness and global assessment are measured.

To improve this situation, future trials in AS should measure health status more comprehensively. Consensus is needed on *what* to measure, i.e., generic or focussed measures in the health status, disease and patient areas of interest. The next step is deciding *how* to measure, i.e., choosing or developing and defining the appropriate measure(s). Several measures with different focus should probably be used simultaneously in a trial.

Finally, to further comprehensive health status measurement in AS we would like to suggest some instruments to complete the existing set. Referring to Figure 1, we would like the addition of more patient reported measures, with varying degrees of focus in the patient, health status, and disease areas. In our view, such measures include the following: (1). The Sickness Impact Profile (SIP) is a generic instrument. It is a self-administered questionnaire which measures physical, mental, emotional and social aspects of function[59], and is applicable across diverse populations and diseases. (2). The AIMS [5] is an example of an instrument more focussed in the disease area. Like the SIP, it is generic in the health status and patient areas of interest. The AIMS is a self-administered questionnaire that assesses physical, emotional and social well being, thus covering the disability and the discomfort dimensions of health status. It has been well validated and is widely used in rheumatoid arthritis and osteoarthritis. After appropriate testing and validation, it may also be useful in AS. (3). Utilities are general measures representing priorities of individual patients. They are, therefore, patient focussed but otherwise generic. Utility measures assess the value or preference a patient attaches to his (her) overall health status. In these measures a patient summarizes the risks and benefits of an intervention into one single value ranging from 0 (death) to 1 (perfect health). This single overall value allows comparison of outcomes across patients between various health care interventions and across different health states or diseases. Preliminary results from a current trial in patients with AS suggest that utility measurement is feasible and reliable.[60] (4). The McMaster Toronto Arthritis Rheumatism Patient Preference Disability questionnaire (MACTAR)[61] and the Problem Elicitation Technique questionnaire (PET)[62] are both patient and disease focussed, but more or less generic in the health status area. Both questionnaires assess patient priorities, allowing each patient to rank items, i.e., activities affected by arthritis which are of high importance to them. These techniques may be more responsive to clinically relevant change than more traditional questionnaires.[61] However, the results of these interviews may also be less easy to generalize.

In summary, to answer clinical questions in AS, we need both generic and focussed measurements to span the whole spectrum of health status. More comprehensive measurement is required to assess health status in AS.

Acknowledgments
The authors wish to acknowledge M.A. Khan and P. Tugwell for their significant contributions.

REFERENCES

1 Kirshner B, Guyatt G: A methodological framework for assessing health indices. J Chron Dis 1985;38:27-36.
2 Fries JF: Toward an understanding of patient outcome measurement. Arthritis Rheum 1983;26:697-704.
3 Fries JF, Spitz P, Kraines RG, Holman HR: Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
4 Deyo RA, Inui TS: Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. Health Serv Res 1984;19:275-90.
5 Meenan RF, Gertman PM, Mason JH: Measuring health status in arthritis: the Arthritis Impact Measurement Scales. Arthritis Rheum 1980;23:146-52.
6 Guyatt GH, Van Zanten SJO, Feeny DH, Patrick DL: Measuring quality of life in clinical trials: a taxonomy and review. Can Med Assoc J 1989;140:1441-8.
7 Coelho Andrade LE, Atra E, Bosi Ferraz M: Sulphasalazine and ankylosing spondylitis: an open pilot study (letter). Clin Exp Rheumatol 1989;7:661-2.
8 Dougados M, Caporal R, Doury P, et al: A double blind crossover placebo controlled trial of ximoprofen in AS (letter). J Rheumatol 1989;16:1167-9.
9 Emery P, Lythgoe S: The effect of acupuncture on ankylosing spondylitis (letter). Br J Rheumatol 1986;25:132-3.
10 Tytman K, Bernacka K, Sierakowski S: D-penicillamine in the therapy of ankylosing spondylitis (letter). Clin Rheumatol 1989;8:419-20.
11 Daltroy LH, Larson MG, Roberts NW, Liang MH: A modification of the Health Assessment Questionnaire for the spondyloarthropathies. J Rheumatol 1990;17:946-50.
12 Domján L, Nemes T, Bálint GP, Tóth Z, Gömör B: A simple method for measuring lateral flexion of the dorsolumbar spine. J Rheumatol 1990;17:663-5.
13 Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B: Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988;15:302-7.
14 Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC: Studies with an enthesis index as a method of clinical assessment in ankylosing spondylitis. Ann Rheum Dis 1987;46:197-202.
15 Astorga G: Double-blind, parallel clinical trial of tenoxicam (Ro 12-0068) versus piroxicam in patients with ankylosing spondylitis. Eur J Rheumatol Inflamm 1987; 9:70-3.
16 Benhamou CL: Large-scale open trials with etodolac (Lodine) in France: an assessment of safety. Rheumatol Int 1990;(suppl)10:29-34.
17 Bernacka K, Tytman K, Sierakowski S: Clinical application of D-penicillamine in ankylosing spondylitis: a 9-month study. Med Interne 1989;27:295-301.
18 Bird HA, Le Gallez P, Astbury C, Looi D, Wright V: A parallel group comparison of tenoxicam and piroxicam in patients with ankylosing spondylitis. Pharmatherapeutica 1986;4:457-62.
19 Busson M: A long-term study of flurbiprofen in rheumatological disorders: III. Other articular conditions. J Int Med Res 1986;14:13-8.
20 Calabro JJ: Efficacy of diclofenac in ankylosing spondylitis. Am J Med 1986;(suppl 4B)80:58-63.
21 Carcassi C, La Nasa G, Perpignano G: A 12-week double-blind study of the efficacy, safety and tolerance of pirazolac b.i.d. compared with indomethacin t.i.d. in patients with ankylosing spondylitis. Drugs Exp Clin Res 1990;16:29-37.
22 Doury P, Roux H.: Isoxicam vs ketoprofen in ankylosing spondylitis. Br J Clin Pharmacol 1986;(suppl 2)22:157-60.

23 Franssen MJ, Gribnau FW, Van de Putte LB: A comparison of diflunisal and phenylbutazone in the treatment of ankylosing spondylitis. Clin Rheumatol 1986;5:210-20.
24 Franssen MJ, van Herwaarden CL, van de Putte LB, Gribnau FW: Lung function in patients with ankylosing spondylitis. A study of the influence of disease activity and treatment with nonsteroidal antiinflammatory drugs. J Rheumatol 1986;13:936-40.
25 Khan MA: A double blind comparison of diclofenac and indomethacin in the treatment of ankylosing spondylitis. J Rheumatol 1987;14:118-23.
26 Lomen PL, Turner LF, Lamborn KR, Brinn EL: Flurbiprofen in the treatment of ankylosing spondylitis. A comparison with phenylbutazone. Am J Med 1986;80:120-6.
27 Lomen PL, Turner LF, Lamborn KR, Brinn EL: Flurbiprofen in the treatment of ankylosing spondylitis. A comparison with indomethacin. Am J Med 1986;80:127-32.
28 Santo JE, Queiroz MV: Oxaprozin versus diclofenac sodium in the treatment of ankylosing spondylitis. J int Med Res 1988;16:150-6.
29 Schwarzer AC, Cohen M, Arnold MH, Kelly D, McNaught P, Brooks PM: Tenoxicam compared with diclofenac in patients with ankylosing spondylitis. Curr Med Res Opin 1990;11:648-53.
30 Corkill MM, Jobanputra P, Gibson T, Macfarlane DG: A controlled trial of sulphasalazine treatment of chronic ankylosing spondylitis: failure to demonstrate a clinical effect. Br J Rheumatol 1990;29:41-5.
31 Davis MJ, Dawes PT, Beswick E, Lewin IV, Stanworth DR: Sulphasalazine therapy in ankylosing spondylitis: its effect on disease activity, immunoglobulin A and the complex immunoglobulin A-alpha-1-antitrypsin. Br J Rheumatol 1989;28:410-3.
32 Dougados M, Boumier P, Amor B: Sulphasalazine in ankylosing spondylitis: a double blind controlled study in 60 patients. Br Med J Clin Res 1986;293:911-4.
33 Feltelius N, Hällgren R: Sulphasalazine in ankylosing spondylitis. Ann Rheum Dis 1986;45:396-9.
34 Fraser SM, Sturrock RD: Evaluation of sulphasalazine in ankylosing spondylitis - an interventional study. Br J Rheumatol 1990;29:37-9.
35 Mielants H, Veys EM, Joos R: Sulphasalazine (Salazopyrin) in the treatment of enterogenic reactive synovitis and ankylosing spondylitis with peripheral arthritis. Clin Rheumatol 1986;5:80-3.
36 Nissilä M, Lehtinen K, Leirisalo-Repo M, Luukkainen R, Mutru O, Yli-Kerttula U: Sulphasalazine in the treatment of ankylosing spondylitis. A twenty-six-week placebo-controlled clinical trial. Arthritis Rheum 1988;31:1111-6.
37 Sadowska-Wróblenska M, Garwolinska H, Maczynska-Rusiniak B: A trial of cyclophosphamide in ankylosing spondylitis with involvement of peripheral joints and high disease activity. Scand J Rheumatol 1986;15:259-64.
38 Zwillich SH, Comer SS, Lee E, Erdman WA, Lipsky PE: Treatment of the seronegative spondylarthropathies with sulphasalazine. J Rheumatol 1988;(suppl 16)16:33-9.
39 Kraag G, Stokes B, Groh J, Helewa A, Goldsmith C: The effects of comprehensive home physiotherapy and supervision on patients with ankylosing spondylitis - a randomized controlled trial. J Rheumatol 1990;17:228-33.
40 Rasmussen JO, Hansen TM: Physical training for patients with ankylosing spondylitis. Arthritis Care Res 1989;2:25-7.
41 Roberts WN, Larson MG, Liang MH, Harrison RA, Barefoot J, Clarke AK: Sensitivity of anthropometric techniques for clinical trials in ankylosing spondylitis. Br J Rheumatol 1989;28:40-5.
42 Tegelberg A, Kopp S: Short-term effect of physical training on temporomandibular joint disorder in individuals with rheumatoid arthritis and ankylosing spondylitis. Acta Odontol Scand 1988;46:49-56.
43 Calin A, Elswood J: The natural history of juvenile-onset ankylosing spondylitis: a 24-year retrospective case-control study. Br J Rheumatol 1988;27:91-3.
44 Calin A, Elswood J: Retrospective analysis of 376 irradiated patients with ankylosing spondylitis and nonirradiated controls. J Rheumatol 1989;16:1443-5.

45 Guillemin F, Briançon S, Pourel J, Gaucher A: Long-term disability and prolonged sick leaves as outcome measurements in ankylosing spondylitis. Possible predictive factors. Arthritis Rheum 1990;33:1001-6.
46 Calin A, Elswood J, Rigg S, Skevington SM: Ankylosing spondylitis - An analytical review of 1500 patients: the changing pattern of disease. J Rheumatol 1988;15:1234-8.
47 Calin A, Elswood J: A prospective nationwide cross-sectional study of NSAID usage in 1331 patients with ankylosing spondylitis. J Rheumatol 1990;17:801-3.
48 Elswood J, Calin A, Berg C, Rogers F: Ankylosing spondylitis. Comparative analysis of Swedish (n=780) and British (n=1500) experience - the National Ankylosing Spondylitis Societies. Scand J Rheumatol 1987;16:437-40.
49 O'Brien BJ, Elswood J, Calin A: Willingness to accept risk in the treatment of rheumatic disease. J Epidemiol Community Health 1990;44:249-52.
50 O'Brien BJ, Elswood J, Calin A: Perception of prescription drug risks: a survey of patients with ankylosing spondylitis. J Rheumatol 1990;17:503-7.
51 Ringsdal VS, Andreasen JJ: Ankylosing spondylitis - experience with a self administered questionnaire: an analytical study. Ann Rheum Dis 1989;48:924-7.
52 Will R, Edmunds L, Elswood J, Calin A: Is there sexual inequality in ankylosing spondylitis? A study of 498 women and 1202 men. J Rheumatol 1990;17:1649-52.
53 Wordsworth BP, Mowat AG: A review of 100 patients with ankylosing spondylitis with particular reference to socio-economic effects. Br J Rheumatol 1986;25:175-80.
54 Rigby AS, Silman AJ: Outcome assessment in clinical trials of ankylosing spondylitis (editorial). Br J Rheumatol 1991;30:321-5.
55 Laurent MR, Buchanan WW, and Bellamy N: Methods of assessment used in ankylosing spondylitis clinical trials: a review. Br J Rheumatol 1991;30:326-9.
56 Bellamy N, Buchanan WW, Esdaile JM, et al: Ankylosing spondylitis antirheumatic drug trials. I. Effects of standardization procedures on observer dependent outcome measures. J Rheumatol 1991;18:1701-8.
57 Bellamy N, Buchanan WW, Esdaile JM, et al: Ankylosing spondylitis anthirheumatic drug trials. II. Tables for calculating sample size for clinical trials. J Rheumatol 1991;18:1709-15.
58 Bellamy N, Buchanan WW, Esdaile JM, et al: Ankylosing spondylitis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs - results of a consensus development (delphi) exercise. J Rheumatol 1991;18:1716-22.
59 Bergner M, Bobbitt RA, Carter WB, Gilson BHS: The sickness impact profile: development and final revision of a health status measure. Med Care 1981;19:787-805.
60 Bakker CH, Rutten M, Van Doorslaer E, Van der Linden S: Health related utility measurement in patients with fibromyalgia or ankylosing spondylitis (abstr). Arthritis Rheum 1990;(suppl):33:140.
61 Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B: The MACTAR patient preference disability questionnaire: an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987;14:446-51.
62 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.

3

HEALTH RELATED UTILITY MEASUREMENT IN RHEUMATOLOGY: AN INTRODUCTION

Carla Bakker, Maureen Rutten[1], Eddy van Doorslaer[2], Kathy Bennett[3], Sjef van der Linden

Department of Internal Medicine, Division of Rheumatology, and the Department of Health Economics[1], University of Limburg, Maastricht; Erasmus University Rotterdam[2], the Netherlands; Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada[3]

Reprinted with permission of Patient Education and Counseling 1993;20:145-52

HEALTH RELATED UTILITY MEASUREMENT IN RHEUMATOLOGY: AN INTRODUCTION

ABSTRACT

Utility measures of health related quality of life are preference values that patients attach to their overall health status. In clinical trials, utility measures summarize both positive and negative effects of an intervention into one single value between 0 (equal to death) and 1 (equal to perfect health). These measures allow for comparison of patient outcomes of different diseases and allow for comparison between various health care interventions.
There are 2 different approaches to utility measurement. The first is to classify patients into categories based on their responses to a number of questions about their functional status, as for instance the Quality of Well Being questionnaire. The second approach is to ask patients to assign a single rating to their overall health by means of rating scale, standard gamble, time trade-off, or willingness to pay. The Quality Adjusted Life Year (QALY) as outcome measure includes both effects in terms of quality and quantity of life. Utilities are used as weights to adjust life years for the quality of life in order to calculate QALYs. Both QALYs and utilities are useful in decision-making regarding appropriate procedures for groups of patients.

INTRODUCTION

Quality of life may be affected by rheumatic diseases. In fact, quality of life is a broad concept of multiple viewpoints and includes all factors that impact upon an individual's life. Health related quality of life includes only those factors that are part of an individual's health, which can be defined according to the World Health Organization as a state of complete physical, mental and social well being.[1]
Recognition of the impact of chronic diseases on the patient and the desirability to evaluate treatment effects has led to the development of instruments that measure quality of life. Health related quality of life instruments commonly used in rheumatology usually assess pain, stiffness and physical mobility. However, these instruments are *specific*, as they aim at a specific disease (e.g. ankylosing spondylitis), or at a specific population of patients (e.g. rheumatic patients). The rationale for specific instruments lies in its potential for increased responsiveness, because

only those aspects of quality of life are included for which a priori change can be expected. The disadvantage is that they cannot be used for comparisons between different patient populations, which is possible with more generic instruments. *Generic* instruments are applicable in a wide variety of populations because they cover a broad spectrum of aspects relative to quality of life. Guyatt et al.[2] distinguishes 2 major subcategories of generic instruments: health profiles and utility measures. Using *health profiles*, scores of separate items are obtained which can be combined in a few subdimensional scores and sometimes into one single index score. An example of a health profile is the Sickness Impact Profile (SIP)[3] which measures physical, mental, emotional and social aspects of function.

UTILITY MEASUREMENT

Utility measures of health related quality of life are single measures of the value or preference that the respondents attach to their overall health status. Respondents can be the general public, health care providers or patients. In health care decision-making it would be very advantageous to have such a single numerical measure that really reflects the value of the overall health improvement. It would be useful in making decisions about treatments for individual patients (clinical decision-analysis) and decisions regarding appropriate procedures and technology for groups of patients (technology assessment).[4]
In this chapter we will discuss patient utilities as a measure of effect in evaluating treatments. In our clinical trials we have chosen to measure patient utilities since they reflect the relative value of different health states to people we believe should benefit from services provided by the health care system.[5] In clinical trials, utility measures can be valuable because patients combine positive and negative effects of an intervention into one single value between 0 (equal to death) and 1 (equal to perfect health). With commonly used generic and specific instruments these positive and negative effects are measured separately, and the investigator has little or no information on the patient's trade-offs among therapeutic improvements and treatment side effects.

Measurement of health related utility
There are 2 approaches to utility measurement.[6] The *first approach* is to classify patients into categories based on their responses to questions about their functional status. In the Quality of Well Being (QWB), formerly called the Index of Well Being, this approach is used. Patients are to complete a questionnaire on their performance within 3 dimensions: mobility, physical and social activity. Each dimension consists of 5, 4 and 5 levels of performance respectively. Patients are thus classified into 1 of the 43 possible combinations of levels. Each combination of levels describes a unique health state. Each health state and a standard list of symptoms and problems have already been valued by the general public and by patients with rheumatoid arthritis by means of a categorical rating from 0, equal to "as bad as dying", to 1, equal to "completely well".[7] These ratings are used as values assigned to each health state into which the patient (responding to the QWB questionnaire), is classified. The values are then modified by the presence or absence of problems and symptoms of the standard list. They are added to establish an overall QWB value ranging from 0 (dead) to 1 (healthy).

The *second approach* to utility measurement is to ask patients directly to assign one value to their overall health. The four methods which are most frequently used to elicit utility values are

rating scale, standard gamble, time trade-off and willingness to pay.
A *rating scale* consists of a line on paper with clearly defined end points or anchors. It requires that the patient identifies the most and least preferred health states to use as anchor, usually labeled as "perfect health" and as "death". Then the patient is asked to place in order of preference his own health state and so-called marker health states on the rating scale between these anchors; such that the intervals between the placements reflect the differences the patient experiences between the health states. A useful visual aid for the rating scale is a large 'thermometer' with a scale from 0 to 100 (Figure 1).

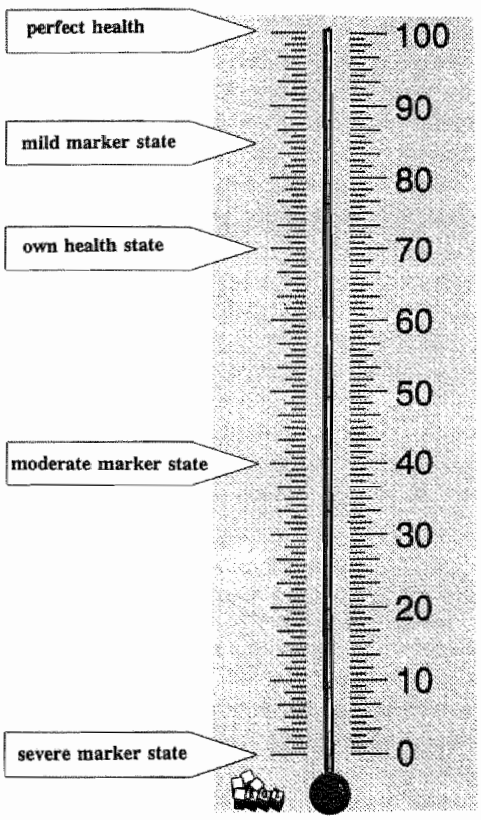


Figure 1. The rating scale, a thermometer

The *standard gamble* technique is based directly on the Von Neumann-Morgenstern utility theory and is the original method of measuring utilities.[8] The standard gamble method consists of paired comparison in which the patient must choose between 2 alternatives. Alternative 1 is a choice with 2 outcomes: either a good outcome, i.e. living in perfect health, with probability p; or a bad outcome, i.e. dying, with probability 1-p. Alternative 2 has one outcome which is

intermediate in desirability between the good and bad outcomes of alternative 1, as for instance the patient's current health state. Probability p is varied until the patient is indifferent to the 2 alternatives. At this point the required utility for the patient's health state has been obtained. It is assumed that patients with a better health accept less risk in order to improve than the more severely affected patients.[9] The standard gamble is supplemented by the use of a probability wheel.

The *time trade-off* method is an implicit technique like the standard gamble, but it does not include risks. Both methods implicitly deduce utilities from the patients' responses to decision situations, whereas in the rating scale method, the preference values are explicitly provided by the patient.[9] The time trade-off is also a paired comparison in which the patient must choose between 2 alternatives. Alternative 1 is to maintain the patient's health state for the rest of their life (time t), while the other alternative is a shorter (time x) but healthy life. Time x is varied until the patient is indifferent to the 2 alternatives, at this point the required preference value for the patient's health state is x/t. It is assumed that the less desirable the patient's health state, the larger the amount of lifetime (in years or months) the patient will trade-off in order to be free from his health state.[9] Usually a visual aid is also used with this technique.

By means of the *willingness to pay* questionnaire, patients are asked how much money they are willing to pay for a hypothetical cure. Thompson reported that rheumatoid arthritis patients were willing to pay 22% of their household income for a complete cure of their arthritis.[10]

These four methods of measuring health related utility are not interchangeable, because they are based on different assumptions and do not all include the risk component.

Patient utilities published in the literature

In the field of rheumatology, only one published randomized controlled drug trial used utility measures. Bombardier et al. reported the results of a multicenter trial in which auranofin (oral gold) was compared with placebo in the treatment of patients with rheumatoid arthritis.[11] Outcome assessment included clinical measures (e.g. number of tender joints and erythrocyte sedimentation rate) and quality of life measures. The last set of measures included arthritis specific instruments (e.g. Health Assessment Questionnaire (HAQ)) as well as generic instruments (e.g. Quality of Well Being Questionnaire (QWB) and Patient Utility Measurement Set (PUMS)). Like the clinical measures, the arthritis specific quality of life scales showed significant improvement of the auranofin group compared to the placebo group. At the same time the auranofin group reported more side effects. By means of an overall assessment as reflected in the utility score (PUMS and QWB), positive and negative effects of treatment were balanced. In the patient's opinion functional improvements were superior to injurious side effects because the PUMS as well as the QWB showed a significant improvement of auranofin in comparison to placebo.

Table 1. Backtranslation of the 6 dimensions of health (Roman numerals) and its levels

I. General daily activities and mobility
Think of limitations caused by tiredness, tightness of the chest or pain while working, doing the housework, shopping, walking, climbing stairs, using public transport, driving a car, cycling, etc.
1) able to perform all daily activities and duties at a normal level of mobility
2) able to perform daily activities, but with some difficulties
3) limited in the performance of daily activities
4) limited considerably in the performance of daily activities
5) unable or hardly able to perform daily activities

II. Personal care
Think of e.g. eating, washing, taking a shower or a bath, going to the toilet, etc.
1) completely capable to perform all self-care activities
2) now and then having difficulty in the performance of self-care activities
3) having difficulty in the performance of self-care activities
4) considerable difficulty in performing self-care activities
5) help needed for all self-care activities

III. Anxieties, frustrations and worries related to the course of the disease
1) no anxieties, no worries, not concerned about the course of the disease
2) normally no anxieties, sometimes concerned about the course of the disease
3) depressed because of the inability to function normally
4) often anxious, often concerned about the course of the disease
5) depressive, unhappy and frustrated

IV. Leisure activities
Think of e.g. going out, practising sports, hobbies, etc.
1) able to participate in all leisure activities without difficulty
2) able to participate in all leisure activities but with some difficulty
3) ability to participate in leisure activities is limited
4) no longer able to participate in any leisure activity which requires a certain degree of physical effort or mobility
5) not able to participate in any leisure activity

V. Pain
1) no pain
2) occasionally pain
3) often mild to moderate pain
4) often severe pain
5) continuously severe pain

VI. Side effects of treatment
Think of e.g. nausea, vomiting and/or diarrhoea, GI upset, skin rash, mouth ulcers.
1) no side effects
2) occasionally mild side effects
3) occasionally moderate - severe side effects
4) often moderate - severe side effects
5) severe side effects

Maastricht Utility Measurement Questionnaire
As an example of an utility measurement instrument we will now describe the Maastricht Utility Measurement Questionnaire. By means of this questionnaire we have elicited utility values from patients with ankylosing spondylitis or fibromyalgia (unpublished data).
The Maastricht Utility Measurement Questionnaire, a Dutch translation and adaptation of the McMaster Utility Measurement Questionnaire[12], will be explained by the next 3 steps: 1) Definition of health, 2) Description of health states and 3) Valuation of health states.

Definition of health
Various authors use various dimensions to define the concept of health. In the Maastricht Utility Measurement Questionnaire health has been defined by 6 dimensions: 1) activities of daily living, 2) self-care functions, 3) emotional functions, 4) leisure activities, 5) pain, and 6) side effects of treatment. Each dimension consists of 5 levels of severity: level 1 reflects the best situation and level 5 the worst (Table 1 shows the backtranslation of the dimensions and its levels of the Maastricht Utility Measurement Questionnaire).

Description of health states
The combination of the levels indicated by the patient in the interview, one for each dimension, was used to define 'patient's own health state'. Marker states were created by the combination of 6 levels, one for each dimension. Perfect health was described by combining the first levels of all 6 dimensions, a severe marker state was described by combining the 5th level of all 6 dimensions. Also a mild and moderate marker state were described by a combination of 6 levels (Table 2 shows the description of the mild marker state). These marker states are valuable during the measurement process, as they encourage the respondent to consider a broad range of possibilities before determining their own health state on the spectrum of possibilities.[4]

Table 2. Description of the mild marker state

Able to perform all daily activities and duties at a normal level of mobility
Completely capable to perform all self-care activities
Normally no anxieties, sometimes concerned about the course of the disease
Ability to participate in leisure activities is limited
Occasionally pain
Occasionally mild side effects

Valuation of health states
The measurement of utilities is performed using rating scale and standard gamble technique. After patients have read the description of the marker states and described their health state, they are firstly asked to rank and value the health states by means of a rating scale, a thermometer with perfect health equal to 100 at the top and a severe marker state equal to 0 at the bottom (Figure 1). In addition, the thermometer gives the patient the opportunity to become familiar with the states and gives the investigator an indication of the ordinal rankings of the

health states and information on the intensity of those preferences. Next, the standard gamble technique is performed with a probability wheel as a prop. In the standard gamble method, the health states are valued under risk, as opposed to under certainty as in the rating scale method.
In patient utility measurement, patients are usually asked how they value their own health state in comparison to perfect health and death (Figure 2a). However, in rheumatic diseases with rather low disease related mortality direct confrontation with the risk of dying may be inappropriate in a preference assessment exercise. Therefore a 2-step utility assessment is suggested. The patients are first asked to value their own health state in comparison to perfect health and the severe marker state (Figure 2b), and then to value the severe marker state in comparison to perfect health and death (Figure 2c), which gives a utility value for the severe marker state. *Utility* values for the patient's **own** health state can be calculated by using the results of these 2 steps of the standard gamble.[13]
At followup visits it is also possible to ask the patients to compare their health state at baseline with their health state at followup. This additional question enables the patient to directly express the change in health related quality of life in a utility value.

To illustrate the method of standard gamble questions that we ask patients, we now present a series of choices. Imagine you are a fibromyalgia patient and to find out how you value your health state at this moment, the first set of alternatives is: Alternative A: 100% chance of living perfectly healthy and 0% chance of living like the severe marker state. Alternative B: living in your current health state (suppose you have fibromyalgia). All health states sustain for the rest of your life. We assume you choose alternative A. The next set is: Alternative A: 10% chance of living perfectly healthy and 90% chance of living like the severe marker state. Alternative B remains the same. We assume your fibromyalgia is not that bad that you take such a big risk and choose alternative B. Then we continue with the next set: Alternative A: 90% chance of living perfectly healthy and 10% chance of living like the severe marker state. Alternative B remains the same. If you feel not that bad, you probably will not take the risk and therefore choose your own health state, the standard gamble stops and this last choice is reported. If you feel your fibromyalgia is that bad that you would take the risk, we continue to the next set: 20% chance of living perfectly healthy and 80% of living like the severe marker state. If you choose alternative B, the next set of chances is 80% chance of living perfectly healthy and 20% of living like the severe marker state. If you choose A, the next set is 30% chance of living perfectly healthy and 70% of living like the severe marker state, et cetera. This is the first step of the 2-step standard gamble question as presented in Figure 2b.

Figure 2a Standard Gamble: Value your own health state in comparison to perfect health and death

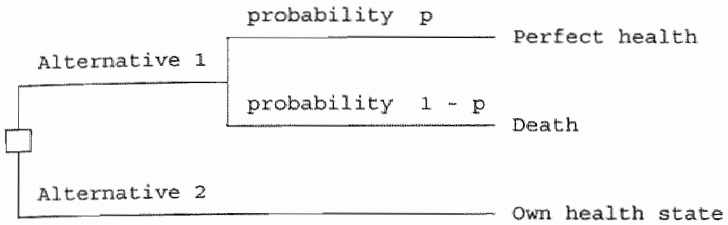


Figure 2b. Two-step Standard Gamble: first step: Value your own health state in comparison to perfect health and the severe marker state

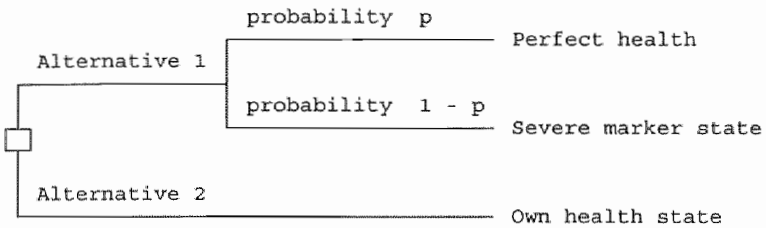
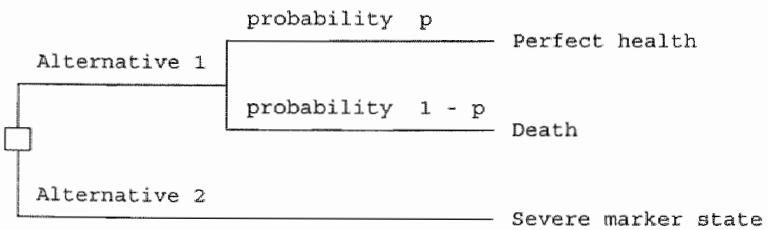


Figure 2c. Two-step Standard Gamble: second step: Value the severe marker state in comparison to perfect health and death



QUALITY ADJUSTED LIFE YEARS

Quality of life measurements are often used in studies of interventions in rheumatic diseases, because these interventions are primarily directed towards the prevention or reduction of morbidity rather than mortality. Most interventions in rheumatology have no effect on survival, but do affect quality of life. These effects can be positive as well as negative. Assigning utilities to intervention-induced changes in quality of life is a way to balance effects of different sizes, both positive and negative, and combine them in an overall summary value. Furthermore, using utilities as outcome measures allows not only a comparison of different interventions in rheumatic diseases, but also a comparison between these interventions and interventions in non-rheumatic diseases which also aim at improving the quality of life.

Besides interventions primarily directed at improving the quality of life, a number of health care interventions affect life expectancy. Moreover, a number of health care interventions do affect both the quality and quantity of life. To allow the comparison of the effectiveness between these various interventions, an additional outcome measure, the so-called Quality Adjusted Life Year (QALY) has been introduced.

A QALY, a concept which was first introduced by Weinstein[14] is a single comprehensive outcome measure that includes effects in terms of both quality of life and survival. Suppose, for example, that the quality of life of an individual patient suffering from ankylosing spondylitis improves from 0.70 to 0.79 by effective drug therapy. This improvement will last for the remaining lifetime of 25 years. Suppose the survival of a patient who has had a transplant heart increases by 2 years at a quality of 0.8 and an additional half year at a quality of 0.6. This patient gains (2 x 0.8) + (0.5 x 0.6) = 1.9 QALY's, whereas the ankylosing spondylitis patient gains 25 x (0.79 - 0.70) = 2.25 QALYs.

In calculating QALYs, the remaining life years are weighted by using a quality-index for the patients' health state during these years. These weights can be elicited by performing utility measurement. Utilities are not the same as QALYs but are used as weights to adjust life years for the quality of life in order to calculate QALYs.[4]

Both utilities and QALYs can be related to costs, resulting in a cost-utility analysis, which is useful to planners and policy makers. QALYs have the potential advantage over utilities that the meaning of costs per unit of utility gained may not have the intuitively appealing meaning as costs per QALY gained. However, there are still a number of problems. Our aim is not to discuss them extensively, but to mention some of them. A major problem is the assumed independence between life years and quality of life which allows the direct multiplication of life years with utilities. However, it is likely that the utility that an ankylosing spondylitis patient assigns to living with moderate ankylosing spondylitis will not only be determined by this particular health state, but also by the number of years this health state is expected to last. For a possible solution to this problem see Mehrez and Gafni.[5,15]

Another problem concerns the comparability of utilities and QALYs across different interventions. A comparison is not allowed when QALYs are not based on the same underlying methods of utility measurement (rating scale, standard gamble, time trade-off) which in turn have to be based on the same underlying dimensions of health. It is shown that different methods of utility measurement do produce essentially different results.[16] And who is to judge which dimensions are the right ones? Dimensions used in various utility measurement instruments so far may not be as sensitive to changes in chronic conditions as they are to

changes achieved by acute care.[17] This raises criticism concerning the consequences of using QALYs for the distribution of health care resources.[18,19]

QALY's are originally developed in the context of cost-utility analysis, thus allowing a broad comparison of interventions across disease categories. Despite a number of problems which still have to be solved, the utility and QALY-approach will become more useful as more and more interventions are analyzed using the same underlying health dimensions and the same underlying methods for obtaining the utilities. By means of utility measurement, value judgement which otherwise implicitly guides decisions about the distribution of health care technologies is now made explicit. This has at least the potential to increase rationality in decision-making.

REFERENCES

1 WHO (1958) The first ten years of the World Health Organization. Geneva: WHO.
2 Guyatt GH, Veldhuyzen Van Zanten SJO, Feeny DH, Patrick DL: Measuring quality of life in clinical trials: a taxonomy and review. Can Med Assoc J 1989;140:1441-8.
3 Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: Development and final revision of a health status measure. Med Care 1981;14:787-805.
4 Torrance GW, Feeny D: Utilities and quality-adjusted life years. Intl J Technology Assess Health Care 1989;5:559-75.
5 Mehrez A, Gafni A: Quality-adjusted life years, utility theory and health-years equivalents. Med Decis Making 1989;9:142-9.
6 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.
7 Balaban DJ, Sagi PC, Goldfarb NI, Nettler S: Weights for scoring the Quality of Well-being instrument among rheumatoid arthritics. A comparison to general population weights. Med Care 1986;24:973-80.
8 von Neumann J, Morgenstern O: Theory of games and economic behavior. Princeton, NJ: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).
9 Torrance GW: Utility approach to measuring health-related quality of life. J Chron Dis 1987;40:593-600.
10 Thompson MS: Willingness to pay and accept risks to cure chronic disease. Am J Public Health 1986;76:392-6.
11 Bombardier C, Ware J, Russell J, Larson M, Chalmers A, Read L: Auranofin therapy and quality of life in patients with rheumatoid arthritis. Am J Med 1986;81:565-78.
12 Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991;(suppl)12:118-28.
13 Drummond MF, Stoddart GL, Torrance GW: Methods for the economic evaluation of health care programmes. Oxford Medical Publications: Oxford, 1987.
14 Weinstein MC, Stasson WB: Foundations of cost-effectiveness analysis for health and medical practices. N Eng J Med 1977;296:716-21.
15 Gafni A: The quality of QALY's (quality-adjusted-life-years): do QALY's measure what they at least intend to measure? Health Policy 1989;13:81-3.
16 Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC: Preferences for Health Outcomes. Comparison of Assessment Methods. Med Decis Making 1984;4:315-29.
17 Donaldson C, Atkinson A, Bond J, Wright K: Should Qualy's be program-specific? J Health Econ 1988;7:239-57.
18 Smith A: Qualms about QALY's. Lancet 1987;i:1134-6.
19 Loomes G, McKenzie L: The use of QALY's in Health Care Decision Making. Soc Sci Med 1989;28:299-308.

4

FEASIBILITY OF UTILITY ASSESSMENT BY RATING SCALE AND STANDARD GAMBLE IN PATIENTS WITH ANKYLOSING SPONDYLITIS OR FIBROMYALGIA

Carla Bakker, Maureen Rutten[1], Eddy van Doorslaer[2], Kathryn Bennett[3], Sjef van der Linden

Department of Internal Medicine, Division of Rheumatology, and the Department of Health Economics[1], University of Limburg, Maastricht; Erasmus University Rotterdam[2], the Netherlands; Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada[3]

Reprinted by permission of the Journal of Rheumatology 1994;21:269-74

FEASIBILITY OF UTILITY ASSESSMENT BY RATING SCALE AND STANDARD GAMBLE IN PATIENTS WITH ANKYLOSING SPONDYLITIS OR FIBROMYALGIA

ABSTRACT

Objective. To assess the feasibility of utility measurement in patients with ankylosing spondylitis (AS) or fibromyalgia (FMS). Patient derived utilities provide overall estimates of the impact of a disease on patient well being.
Methods. The Maastricht Utility Measurement Questionnaire was applied cross sectionally to 57 outpatients with AS and 86 outpatients with FMS. By means of rating scale and standard gamble techniques, patients were asked to value their own health state.
Results. All 143 patients completed the interview. Patients with AS valued their personal health state on the rating scale (0-100) considerably higher than patients with FMS (AS: 69 and FMS: 54). Standard gamble utility values (0-1), however, were about the same at a higher level (AS: 0.86 and FM: 0.83). Four weeks test-retest reliability was examined in 15 patients with FMS. The intraclass correlation coefficient of the utility score for the patient's own health state was 0.56 for the rating scale and 0.66 for the standard gamble technique.
Conclusion. Feasibility of the Maastricht Utility Measurement Questionnaire was generally satisfactory in both patient groups. Utility values obtained by rating scale and standard gamble technique differed considerably. Our data support the view that utility measurement is sensitive to the method chosen to elicit patient well being.

INTRODUCTION

Chronic rheumatic diseases have a major impact on the quality of life rather than the length of life. Nowadays, clinical decisions and health care programs frequently take quality of life into account. This has led to the development of a whole array of instruments to measure well being or quality of life.
Currently, interest is growing in evaluating patient preferences for alternative therapeutic interventions.[1]
Utility measures of health related quality of life are generic measures of the value or preference that patients attach to their overall health status or well being. Utility measures can be used for

comparisons across different patient populations. In clinical trials, utility measures are valuable because patients have to combine the positive and negative effects of an intervention into one single value. In contrast, with commonly used generic and specific instruments these benefits and risks are measured separately, and the investigator has little or no information on the relative value to patients of therapeutic improvements and treatment side effects.[2] As an additional advantage utility measurement provides one overall value which allows the patient outcomes of different diseases or resulting from various health care interventions to be compared.

Recently, utility measurement has been used in the auranofin trial in rheumatoid arthritis (RA)[3], and also in the evaluation of patients' willingness to accept risk in the drug treatment of ankylosing spondylitis (AS) in comparison to RA.[4]

We elicited utility values in 2 randomized controlled trials on patients with AS or fibromyalgia (FMS) to assess the feasibility and reliability of utility measurement in these rheumatic diseases. We compared 2 different techniques to obtain patients' utilities: the rating scale and standard gamble technique.[5]

MATERIALS AND METHODS

Study populations

AS. A total of 72 patients with AS from the outpatient departments of rheumatology of the Maastricht University Hospital, the De Wever Hospital in Heerlen and Sittard's Maasland Hospital completed a randomized controlled trial comparing 2 nonsteroidal antiinflammatory drugs (NSAID). Eligibility criteria were as follows: age 21-55 years, modified New York criteria[6], exacerbation of AS symptoms 3-7 days after stopping NSAID treatment, and written informed consent. Altogether 57 (79%) of the 72 patients were willing to take part in our study dealing with utility measurement.

FMS. A study on the overall therapeutic effect of fitness training and biofeedback[7,8] was about to start at the outpatient rheumatology department of the Maastricht University Hospital. Altogether 86 patients with FMS met the eligibility criteria: female sex, age 18-60 years, fulfillment of the criteria of Wolfe, *et al.*[9] Patients were excluded if they had a high depression on 6 scales of the symptom check-list (SCL-90).[10,11] The study was restricted to female patients for practical reasons. All these patients agreed to participate in the utility study.

To assess the reliability of the utility measurements 15 of the 86 patients with FMS who were to receive no specific therapeutic intervention (controls) were included in the study. Characteristics of the patients with AS and FMS are shown in Table 1.

Table 1. Characteristics of patients with AS and FMS

	AS		FMS	
number of patients	57		86	
% female	35		100	
mean age (SD)	38	(7.6)	44	(8.4)
mean duration of disease: yrs (SD)	16	(7.9)	12	(9.8)
% using NSAID	100		20	

Utility measures

To elicit utility values the Maastricht Utility Measurement Questionnaire (MUMQ), a Dutch translation and adaptation of the McMaster Utility Measurement Questionnaire[12], was administered by trained interviewers. Both patient groups were asked to value their current state of health (i.e., to indicate utility). In the AS trial, the MUMQ was applied at the end of the NSAID trial. This implies that the results refer to the patients' state of health after the intervention, whereas in the FMS trial it was applied just before the beginning of the study.
The MUMQ will now be explained in 3 steps: conceptualization of health, description of health states, and valuation of health states.

Conceptualization of Health

Health was defined by 6 dimensions: activities of daily living, self-care functions, emotions, leisure activities, pain, and side effects of treatment. A disease may affect one or more of these dimensions. Each dimension consists of 5 levels of severity: level 1 reflecting the best situation and level 5 the worst (Table 2). To insure that the dimensions were culturally appropriate we adapted the original McMaster dimensions to our setting. The Dutch translation of the 3rd dimension did not completely cover the term 'discomfort' (Table 2). We adapted "occasional moderate side effects" in level 3 of the 6th original McMaster dimension to "occasional moderate to severe side effects". We did this to better distinguish from level 2 "occasional mild side effects". We also adapted the example used in the first original McMaster dimension: "stopped taking the bus to work" was replaced by "cycling," a typically Dutch activity.

Table 2. Backtranslation of the 6 dimensions of health (Roman numerals) and their levels

I. General daily activities and mobility

Think of limitations caused by tiredness, tightness of the chest or pain while working, housework, shopping, walking, climbing stairs, using public transport, driving a car, cycling, etc.

1) able to perform all daily activities and duties at a normal level of mobility
2) able to perform daily activities, but with some difficulties
3) limited in the performance of daily activities
4) limited considerably in the performance of daily activities
5) unable or hardly able to perform daily activities

II. Personal care

Think of eating, washing, taking a shower or a bath, going to the toilet, etc.

1) completely able to perform all self-care activities
2) now and then having difficulty in the performance of self-care activities
3) having difficulty in the performance of self-care activities
4) considerable difficulty in performing self-care activities
5) help needed for all self-care activities

III. Anxieties, frustrations and worries related to the course of the disease

1) no anxieties, no worries, not concerned about the course of the disease
2) normally no anxieties, sometimes concerned about the course of the disease
3) depressed because of the inability to function normally
4) often anxious, often concerned about the course of the disease
5) depressive, unhappy and frustrated

IV. Leisure activities

Think of going out, practicing sports, hobbies, etc.

1) able to participate in all leisure activities without difficulty
2) able to participate in all leisure activities but with some difficulty
3) ability to participate in leisure activities is limited
4) no longer able to participate in any leisure activity which requires a certain degree of physical effort or mobility
5) not able to participate in any leisure activity

V. Pain

1) no pain
2) occasional pain
3) often mild to moderate pain
4) often severe pain
5) continuous severe pain

VI. Side effects of treatment

E.g. nausea, vomiting and/or diarrhoea, gastro-intestinal upset, skin rash, mouth ulcers.

1) no side effects
2) occasional mild side effects
3) occasional moderate - severe side effects
4) often moderate - severe side effects
5) severe side effects

Description of Health States
The combination of the levels indicated by the patient in the interview, one for each dimension, was used to define "patient's own health state". Marker states were created as follows: perfect health was the combination of the first levels of all 6 dimensions, a severe disease marker state was described as the combination of all 5th levels of the 6 dimensions. Mild and moderate marker states were also created (Table 3 shows the description of the mild marker state). The same marker states were used in all patients with AS and FMS. The marker states to describe health states were developed using consensus techniques by a panel consisting of rheumatologists and methodologists.[12]
The mild, moderate and severe disease marker states were presented to the patients as examples of a mild, moderate or severe AS or FMS, respectively. These marker states present the respondent with a broad range of health states before they determine their own health state from the whole spectrum of possibilities.[13]

Table 3. Description of the mild marker state

Able to perform all daily activities and duties at a normal level of mobility
Completely able to perform all self-care activities
Normally no anxieties, sometimes concerned about the course of the disease
Ability to participate in leisure activities is limited
Occasional pain
Occasional mild side effects

Valuation of Health States
After the patients read the description of the marker states and defined their own health state, they were first asked to rank and value these states by means of a *rating scale*, a thermometer with perfect health equal to 100 at the top and the severe marker state equal to 0 at the bottom. The rating scale gives the investigator the ordinal rankings of the health states and information on the strength of those preferences.
Thereafter, the *standard gamble* technique is performed with a probability wheel as a prop.[14] The standard gamble is directly based on the Von Neumann-Morgenstern utility theory and is the original method of measuring utilities.[15] In the standard gamble method, health states are valued under the assumption of risk, as opposed to the rating scale method where risk is not included in the measurement process. The standard gamble method directly delivers utility values. Torrance has suggested that rating scale values be corrected by a power curve approach to approximate utility values.[5]
In utility measurements, patients are usually asked how they value their own health state in comparison to perfect health and death (Figure 1a). However, in rheumatic diseases with rather low disease related mortality direct confrontation with the risk of dying seemed inappropriate in a preference assessment exercise. Therefore a 2-step utility assessment was performed. The patients are first asked to value their own health state in comparison to perfect health and the severe marker state (Figure 1b) rather than death. Next, they are asked to value the severe marker state in comparison to perfect health and death (Figure 1c), which gives a utility value for the severe marker state on the standard 0-1, death-healthy scale. The first step provides a

standard gamble *score* which must be converted into a *utility value* for the patient's own health state using the utility of the severe marker state.[16] Negative utility values for the severe marker state, indicating that this state was considered worse than death, were recoded to zero in this analysis.

Figure 1a. Standard Gamble: Value your own health state in comparison to perfect health and death

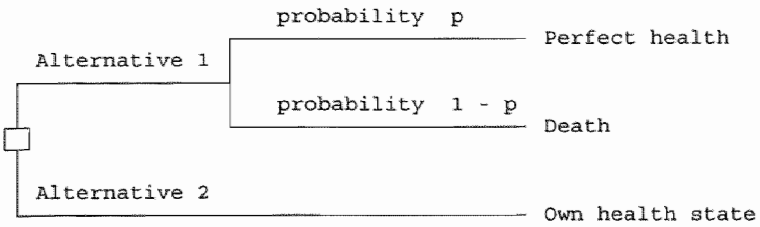


Figure 1b. Two-step Standard Gamble: first step: Value your own health state in comparison to perfect health and the severe marker state

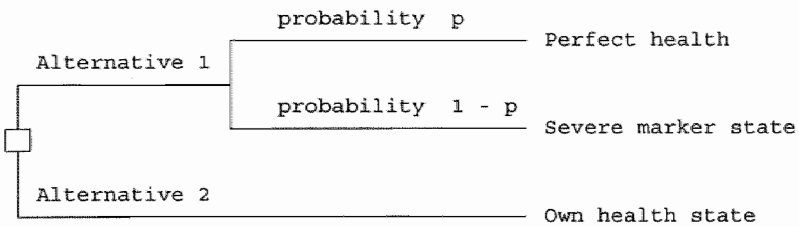
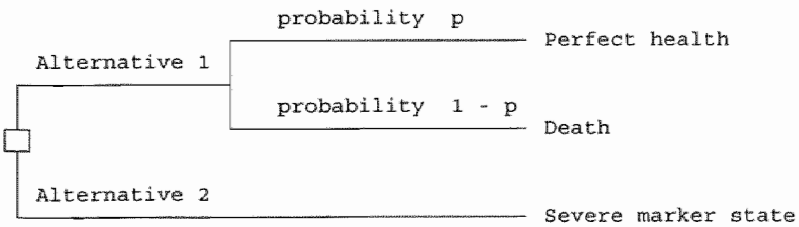


Figure 1c. Two-step Standard Gamble: second step: Value the severe marker state in comparison to perfect health and death



Other outcome measures
In addition to utility measurements, global health, functioning and pain were assessed. Global health was indicated on a 5 point scale with 0 = "very good", 1 = "good", 2 = "moderate", 3 = "poor", and 4 = "bad," whereas functioning was assessed by the Sickness Impact Profile (SIP) questionnaire.[17] The overall SIP percent score of dysfunction was reported. Pain was reported on a 100 mm visual analog scale with 0 = "no pain" and 100 = "most imaginable pain".

Statistics
Test-retest reliability of the MUMQ was tested by intraclass correlation coefficient in a control group of 15 patients with FMS.

RESULTS

Feasibility
We assessed the proportion of patients who completed the questionnaire successfully and measured the duration of the interview. No interviews were broken off. Four (2.8%) of all 143 patients gave inconsistent answers (Table 4): one on the rating scale and 3 on the standard gamble. One patient placed the moderate marker state higher on the rating scale than the mild marker state, one patient preferred her own health state to a 100% chance of being perfectly healthy, and 2 patients preferred a marker state even where they had a 100% chance of being perfectly healthy. These inconsistent answers were excluded from the analysis.
The duration of the interview was almost the same as in patients with AS and in patients with FMS. The mean duration for the patients with AS was 12 min (SD, 3.6) for the rating scale and 14 min (SD, 5.4) for the standard gamble; for the patients with FMS it was 11 min (SD, 2.9) for the rating scale and 12.5 min (SD, 3.8) for the standard gamble.

Reliability
The control group of 15 patients with FMS was retested after 1 month. No specific intervention had been given to these control patients during this period. Marker states are assumed to be constant over time. The intraclass correlation coefficients of the utility scores were 0.56 and 0.66 for the patient's own health state on the rating scale and standard gamble, respectively, 0.67 and 0.74 for the mild marker state on the rating scale and standard gamble, respectively, and 0.94 for the severe marker state (standard gamble).

Valuation of Health States
The mean values for the mild and moderate marker states on the rating scale did not differ very much between AS and FMS (Table 4). The value for the patient's own health state on the rating scale was 69 for the patients with AS who had just completed a NSAID trial and 54 for the selected group of patients with FMS who were about to receive a new therapeutic intervention.
By standard gamble the mean utilities for the marker states and for the patient's own health state were higher than by rating scale. Also, the mean standard gamble utility value for patient's own health state did not differ very much between AS (0.86) and FMS (0.83) (Table 4).
The sizes of the SD for the marker states and the patients' own health state did not differ clearly between AS and FMS for both rating scale and standard gamble.
In the standard gamble more patients with FMS (19%) than AS patients (9%) were not prepared

to accept the lowest possible risk (10%) of death to obtain perfect health. They preferred the certainty of continuing life in the severe disease state. This in turn resulted in a high utility (0.95) for the severe marker state for these patients. Altogether 43% of the patients with AS and 28% of the patients with FMS preferred death to the severe marker disease state.

Table 4. Maastricht Utility Measurement Questionnaire: comparison of patients with AS and FMS

		AS		FMS	
Number of interviews		57		86	
Broken off		0		0	
Inconsistent answers		0		4	
Marker States Mean (SD)					
Rating Scale (0-100)	mild	71	(14)	73	(14)
	moderate	23	(15)	27	(13)
Standard Gamble Utility (0-1)	mild	0.86	(0.21)	0.85	(0.20)
	severe	0.34	(0.37)	0.45	(0.38)
Own Health State Mean (SD)					
Rating Scale (0-100)		69	(22)	54	(18)
Standard Gamble Utility (0-1)		0.86	(0.21)	0.83	(0.19)

Other outcome measures

Patients with AS scored higher in global health than patients with FMS (Table 5). This last group also reported more dysfunction and pain than patients with AS on SIP questionnaire and pain VAS. Generally, the rating scale results agreed better with these other outcomes than the standard gamble utility values.

Table 5. Utility values, global health, SIP overall percent score and pain* among patients with AS and FMS

	AS		FMS	
Rating scale (0-100)	69	(22)	54	(18)
Standard gamble utility (0-1)	0.86	(0.21)	0.83	(0.19)
Global health (0-4)	1.47	(1.02)	2.87	(0.72)
SIP overall % score	4.2	(5.0)	14.0	(8.6)
Pain (0-100)	30.1	(21.0)	59.5	(19.4)

* See text for global health scale and VAS.

DISCUSSION

Patient derived utilities provide overall estimates of the impact of a disease on patients' well being. By the MUMQ we assessed utilities of 2 groups of patients with AS and FMS with 2 different techniques. We applied both the rating scale and the standard gamble method.
The feasibility of applying the questionnaire was generally satisfactory in both patient groups. It should be realized that 21% of the patients with AS did not agree to spend the amount of time and energy necessary to complete the interview, but this occurred after a demanding one year trial with many followup visits. Although both the rating scale and standard gamble have been said to be difficult techniques for obtaining utilities[14], no interviews were broken off. Reliability based on measurements among 15 patients with FMS appeared satisfactory.
As stated before, our outpatient groups can not be regarded as representative for other patients with these diseases. Therefore, the obtained utility values may not be generalizable to other patients with AS and FMS. For example, the patients with AS had just completed a drug trial and received at the end of the trial the MUMQ addressing their current (posttrial) state of health. On the other hand, patients with FMS were just about to start a therapeutic intervention on the effects of fitness or biofeedback training.
Both patients with AS and FMS valued their own health state higher on the standard gamble than on the rating scale. This might be a reflection of the fact that the rating scale does not incorporate a risk of a not preferred outcome.[14,18]
Using the standard gamble scores for the severe marker state (instead of death as the anchor) in calculating indirectly the standard gamble utilities for the patient's own health state has certain consequences. Numerically impressive differences in standard gamble scores for the patient's own health may give rise to only small differences in the utility value for own health (Figure 2). The technique we used required 2 steps. In the first step the patient's own health state was measured on a scale ranging from perfect health to severe marker state. In step 2, the severe marker state was measured on a scale ranging from perfect health to death. This was done to spare patients with chronic diseases from having to make decisions that included a risk of death. One might imagine that this 2-step indirect measurement approach provides different utility values for the patient's own health state than utility values obtained by asking the patient to value their own health state **directly** in comparison to perfect health and death. This point should be validated by future research.

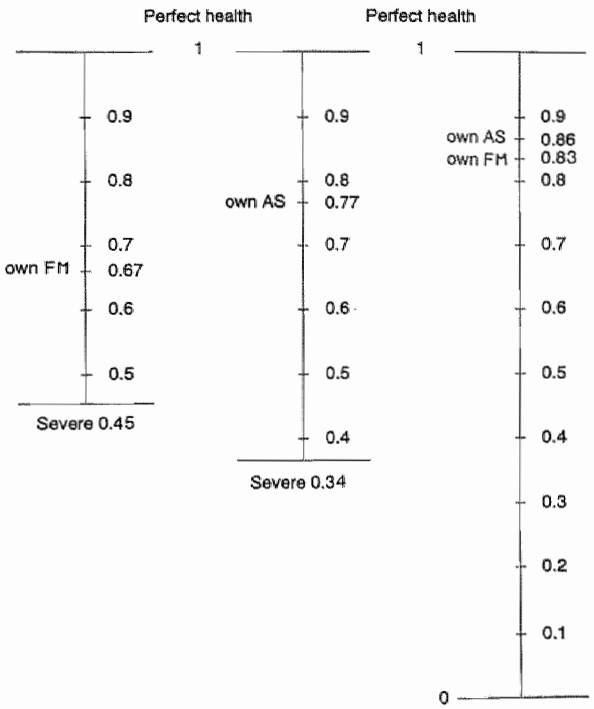


Figure 2. Calculation of utilities *

* Different valuations of the benchmark, the severe marker state, results in scales as shown above. In calculating **utilities** for the patient's own health state, the standard gamble **scores** for their own health state are brought together on the same 0-1 scale (1 = perfect health and 0 = death). The formula is: Uo = p + (1-p)*Us
Uo = utility of the patient's own health state, Us = utility of the severe marker state and p = probability level at which the patient is indifferent between the 2 alternatives.

The utility of the severe marker state provides a benchmark in calculating the utilities for the other health states. The mean value of the severe marker state differed between patients with AS and FMS (Table 4). There are 2 possible causes for this difference in the valuation of the severe marker state. Firstly, the severe marker state is presented to both groups of patients as a severe example of their disease. So, patients with AS value a patient with severe AS, whereas patients with FMS value a patient with severe FMS. This means that patients with AS and FMS may refer to a different health state although the description of the health state presented to

them is the same. Secondly, risk attitudes may differ between patients who have different diseases. The percentage of patients who preferred death to the severe marker state indeed differed between AS (9%) and FMS (19%).
For both diseases negative standard gamble scores for the severe marker state have been recoded to zero, which causes the calculated utilities to be higher than they would be if a negative value was assigned when the severe state was considered to be worse than death. Another way to deal with this problem is to ask the patients an extra standard gamble question in which they can indicate how negative the severe marker state is in their opinion.[19]
A striking finding from our study is the fact that utility values obtained by rating scale and standard gamble technique differed considerably. This has important consequences. Utility values are used to adjust qualitatively the length of survival (Quality Adjusted Life Years - QALY). In cost-effectiveness studies the costs / QALY are calculated. In a recent study comparing 6 different instruments of assessing well being, large variation was found with considerable discrepancies at the individual level. It was concluded that the substantial variability in patients' stated quality of life may preclude the use of a single method to analyze the cost-effectiveness of a health care program.[18] Our data strongly support these findings indicating that utility measurement is sensitive to the method chosen to elicit patient well being. Interinstrument variation is less critical in the followup of individual patients once these instruments have shown to be valid in other respects. Indeed, further research dealing with utilities should focus on aspects such as validity and sensitivity to clinically important change. Also, the validity of different methods of utility measurement requires further study. The real challenge is to come up with a suitable and acceptable definition of the gold standard for utility measurement.
In summary, our results indicate that in the context of a trial it is generally feasible to apply the MUMQ to obtain utilities. Our data support the view that utility measurement is sensitive to the method chosen to elicit patient well being.

Acknowledgment
The authors acknowledge P. Bolwijn MSc, and Drs P. Seys and M. van Santen-Hoeufft for their help in recruiting the patients, and M. Aarts, MD, and R. de Bie, MSc, for their help *in* interviewing the patients.

REFERENCES

1 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.

2 Feeny D, Labelle R, Torrance GW: Integrating economic evaluations and quality of life assessments. In: Spilker B, ed. Quality of life assessments in clinical trials. New York: Raven Press, 1990.

3 Bombardier C, Ware J, Russell J, Larson M, Chalmers A, Read, L: Auranofin therapy and quality of life in patients with rheumatoid arthritis. Am J Med 1986;81:565-78.

4 O'Brien BJ, Elswood J, Calin A: Willingness to accept risk in the treatment of rheumatic disease. J Epidemiol Community Health 1990;44:249-52.

5 Torrance GW: Utility approach to measuring health-related quality of life. *J Chron Dis* 1987;40:593-600.

6 Linden van der S, Valkenburg HA, Cats A: Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.

7 Ferraccioli G, Ghirelli L, Scita F, et al: EMG-Biofeedback training in fibromyalgia syndrome. J Rheumatol 1987;14:820-5.

8 McCain G, Bell D, Mai F, Halliday P: A controlled study of the effects of a supervised cardiovascular fitness training program on the manifestations of primary fibromyalgia. Arthritis Rheum 1988;31:1135-41.

9 Wolfe F, Smythe HA, Yunus MB, et al: The American College of Rheumatology 1990. Criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. Arthritis and Rheum 1990;33:160-72.

10 Arrindell WA, Ettema JHM: Handleiding bij een multidimensionele psychopathologie-indicator. Lisse: Swets & Zeitlinger, 1986.

11 Derogatis LR: SCL-90: Administration, scoring and procedures manual-I for the r(evised) version. Baltimore: John Hopkins University School of Medicine, Clinical Psychometrics Research Unit, 1977.

12 Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991; (suppl) 12:118-28.

13 Torrance GW, Feeny D: Utilities and quality-adjusted life years. Intl J Technology Assess Health Care 1989;5: 559-75.

14 Torrance GW: Social preferences for health states: an empirical evaluation of three measurement techniques. Socioecon Planning Sci 1976;10:129-36.

15 von Neumann J, Morgenstern O: Theory of games and economic behavior. Princeton: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).

16 Torrance GW: Measurement of health-state utilities for economic appraisal: A review. J Health Econ 1986;5:1-30.

17 Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.

18 Hornberger JC, Redelmeier DA, Petersen J: Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. J Clin Epidemiol 1992;45:505-12.

19 Torrance GW, Boyle MH, Horwood SP: Application of multi-attribute utility theory to measure social preferences for health states. Operations Res 1982;30:1043-69.

5

PATIENT UTILITIES IN ANKYLOSING SPONDYLITIS AND THE ASSOCIATION WITH OTHER OUTCOME MEASURES

Carla Bakker, Maureen Rutten[1], Alita Hidding, Eddy van Doorslaer[2], Kathryn Bennett[3] and Sjef van der Linden

Department of Internal Medicine, Division of Rheumatology, and the Department of Health Economics[1], University of Limburg, Maastricht; Erasmus University Rotterdam[2], the Netherlands; Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada[3]

Reprinted by permission of the Journal of Reumatology 1994;21:1298-1304

PATIENT UTILITIES IN ANKYLOSING SPONDYLITIS AND THE ASSOCIATION WITH OTHER OUTCOME MEASURES

ABSTRACT

Objective. To compare in patients with ankylosing spondylitis (AS) utilities derived by rating scale and standard gamble method, to relate these values to other outcome measures, and to assess the sensitivity to change of utilities relative to changes in other outcomes.

Methods. Patients with AS were randomly allocated to either weekly sessions of supervised group physical therapy for a period of 9 months or daily exercises at home. Analysis was restricted to the 59 patients who completed the Maastricht Utility Measurement Questionnaire (MUMQ) at baseline and after 9 months' followup and who were seen by the same interviewer. Reliability was assessed by intraclass correlation coefficient and change scores for marker states of disease. Construct validity was evaluated by correlation and multiple regression of baseline values with a variety of disease outcomes (pain and stiffness, patient's and physician's global assessment, Sickness Impact Profile, Arthritis Impact Measurement Scale, Health Assessment Questionnaire for the Spondyloarthropathies, functional, articular, and enthesis indices and spinal mobility measures). Sensitivity to change was assessed against changes in these outcome measures at followup.

Results. The test-retest intraclass correlation coefficients for patient utilities were 0.95 (rating scale) and 0.79 (standard gamble), and for the marker state of mild disease 0.70 (rating scale) and 0.77 (standard gamble). A multiple regression analysis with the *baseline* rating scale or standard gamble utilities as dependent variable showed that patient's global assessment explained 59 and 11% of the total variance respectively. By multiple regression analysis 10% of the variance of *change* in rating scale utilities was explained by *changes* of patient's global assessment. In contrast, variance in change in standard gamble utilities was not explained by changes in other disease outcomes.

Conclusion. Findings obtained by rating scale and standard gamble differ considerably. Standard gamble utilities seem to address different aspects of health status than do rating scale utilities and more traditional outcomes. Utility measurement is sensitive to the method chosen to elicit patient well being.

INTRODUCTION

Utilities are generic health related quality of life measures assessing the value or preference that patients attach to their overall health status: i.e., patients have to integrate all positive and negative effects of their disease and its treatment into one single value. By contrast, in disease specific instruments the benefits and risks are measured separately, and the investigator has little or no information on the relative weights patients assign to therapeutic improvements and disadvantages, such as side effects of drugs or the number of visits to clinics.[1]
Recently, utility measurements have been successfully used in the evaluation of therapeutic interventions for patients with rheumatoid arthritis.[2] Also utility measurements have been applied to patients with ankylosing spondylitis (AS) to assess their willingness to accept risk in drug treatment.[3]
In a randomized controlled trial in AS we found supervised group physical therapy to be superior to exercises at home in improving thoracolumbar mobility, fitness, and global health.[4] During this trial we elicited also utility values both by rating scale and standard gamble method.

We report aspects of validity of utility measurement. We address in particular reliability, construct validity and sensitivity to change related to improvements in other outcomes of AS.

PATIENTS AND METHODS

Study population
Altogether, 163 patients with AS gave written informed consent to participate in the study. After checking the inclusion (modified New York criteria)[5] and exclusion criteria by one rheumatologist 144 patients remained.[4] Patients were randomized to unsupervised daily individualized exercises at home or the same plus weekly group physical therapy for 9 months. Group therapy sessions consisted of 1 h of physical training, followed by 1 h of sporting activities and 1 h of hydrotherapy.[4] Our report is confined to the 59 patients with AS, to whom the Maastricht Utility Measurement Questionnaire (MUMQ) was applied by the same interviewer at baseline and after 9 months' followup. The 2 interviewers were blinded as to the intervention. Demographic and clinical characteristics for the 59 patients with AS are shown in Table 1.

Table 1. Baseline characteristics for the 59 patients with AS

Age (years)		
mean (SD)		44 (10.4)
Duration of disease (years)		
mean (SD)		5.9 (5.9)
Sociodemographic characteristics %		
Male		71
Married		80
Employed		61
Education* level:	High	16
	Middle	43
	Low	41

* Years of education (including primary school);
high > 15 years; middle 10-15 years; low < 10 years

Utility measures

The MUMQ, a Dutch translation and adapted version of the McMaster Utility Measurement Questionnaire[6], has been described.[7] Briefly, patients were first asked to define their own health state by indicating the level of health on 6 dimensions, i.e., physical state and mobility, self-care, emotions, leisure activities, pain, and side effects of treatment (Table 2). Marker states (perfect health, mild AS and severe AS) were created by combining 6 levels, one for each dimension. Patients were asked to value the provided marker states and their own health status, using both the rating scale and standard gamble method.

The *rating scale* is a numerical scale and looks like a thermometer with "perfect health" equal to 100 at the top and the marker state of "severe disease" equal to 0 at the bottom.

The *standard gamble* is performed with a probability wheel as a prop.[8] The standard gamble is the original method of measuring utilities and directly based on the Von Neumann-Morgenstern utility theory.[9] In the standard gamble method health states are valued under the assumption of risk, as opposed to the rating scale method where risk is not included in the measurement process. A 2-step standard gamble utility assessment was performed: patients are first asked to value their own health status in comparison to a gamble with probability (p) to gain perfect health and a probability (1-p) to attain the *severe* marker state (Figure 1a). Next, they are asked to value the severe marker state in comparison to a gamble with probability (p) to gain perfect health and a probability (1-p) to die (Figure 1b). Probability (p) is systematically varied until the patient is indifferent between the 2 alternatives. In our study (p) was varied with steps of 10% [(p)/(1-p): 100/0, 10/90, 90/10, 20/80, 80/20, 30/70 etc.]. When indifference is reached, utilities for the health states in alternative 2 are calculated by: $U = pU_{best} + (1\text{-}p)U_{worst}$, where U_{best} is the utility of the best outcome of the gamble (perfect health, and its utility is equal to 1 by definition) and U_{worst} is the utility of the worst outcome of the gamble (severe marker state in step 1; and death in step 2, and the utility of death is 0 by definition). The 2nd step is measured between perfect health and death and therefore directly provides a utility value for the severe marker state. The first step provides a standard gamble *score*, which has to be converted into a

utility value for the patient's **own** health status, using the utility of the severe marker state.[10] Negative utility values for the severe marker state, indicating that a patient considered this state worse than death, were recoded to zero in the analysis.

Figure 1a. Two-step Standard Gamble: first step: Value your own health status in comparison to perfect health and the marker state of severe disease

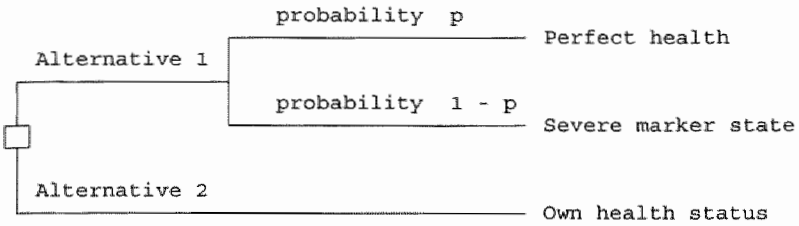
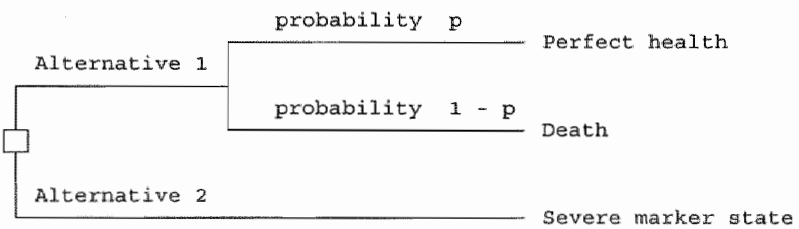


Figure 1b. Two-step Standard Gamble: second step: Value the marker state of severe disease in comparison to perfect health and death



As other outcomes the following health status measures were applied: the Sickness Impact Profile (SIP)[11]; the Dutch Arthritis Impact Measurement Scale (Dutch-AIMS)[12]; Health Assessment Questionnaire for the Spondyloarthropathies (HAQ-S)[13]; and the functional index.[14] The following assessments were also made: the articular index[14]; enthesis index[15]; spinal mobility by chest expansion, cervical rotation, and thoracolumbar flexion and extension[16]; physical fitness (or aerobic power) by bicycle ergometer.[4] Pain and stiffness were indicated by the patient on a horizontal 100 mm visual analogue scale (VAS) with 0 equal to "no pain or stiffness" and 10 equal to "worst pain or stiffness I can imagine." Physician's global assessment was assessed at baseline on a 5 point scale with 1 equal to "low disease activity" and 5 equal to "high disease activity." Patient's global assessment was assessed by asking the patient to describe his or her perceived change in general functioning after treatment on a 10 cm horizontal VAS (-5 = maximum worsening, 0 = no change, +5 = maximum improvement).

Statistical methods

Reliability was tested by intraclass correlation coefficient among 14 participants to whom the MUMQ was again applied after 1 week. Also after 9 months' followup stability of marker states was tested by Wilcoxon signed rank test among the patients with AS.

Construct validity of a measure indicates whether it changes in proportion to how a patient changes clinically.[17] Construct validity of MUMQ was tested 1-sidedly by the Spearman correlation coefficient between *baseline* utility values, and *baseline* scores for spinal mobility (chest expansion, spinal flexion/extension, cervical rotation), physical fitness, pain, stiffness, patient's and physician's global assessment, scores for the functional, articular and enthesis indices, and self-addressed questionnaires such as SIP, AIMS, and HAQ-S. In addition, multiple regression analysis (stepwise forward) was performed with all these measures and also disease duration, age, marital status, education and sex as independent variables and rating scale or standard gamble utilities as dependent variable. Independent variables with skewed distributions were analyzed as ln(var+1), the natural logarithm of one plus the variable.[18,19]

Discriminant validity indicates whether a measure is sensitive to change, i.e., whether it can detect important clinical changes in disease activity over time.[17] In our study discriminant validity was tested in 2 ways. First, we calculated Spearman correlation coefficients (1-sided testing) between the *changes* in utility values and the *changes* in other outcome measures. Secondly, we performed multiple regression analyses (stepwise forward) with *changes* in rating scale or standard gamble utilities as dependent variable and the trial intervention and *changes* in other health status outcome measures as independent variables. Both dependent and independent variables in the multiple regression analysis were transformed by ln(var+constant), the natural logarithm of a constant just below the minimum of (change in) the variable plus the (change in) the variable.

Table 2. Back translation of the 6 dimensions of health (Roman numerals) and their levels

I.	**Physical state and mobility**
1)	able to perform all duties at home and at work without difficulty
2)	able to perform all duties at home and at work, but with some difficulty
3)	not able to perform some duties at home and/or at work
4)	not able to perform many duties at home and/or at work
5)	not able to perform any duties at home and/or at work
II.	**Self care**
1)	able to perform all self-care activities without any help
2)	able to perform all self-care activities, though not without effort
3)	not able to perform some self-care activities without help
4)	not able to perform many self-care activities without help
5)	not able to perform any self-care activities without help
III.	**Emotions**
1)	I do not worry about my illness
2)	I seldom worry about my illness
3)	occasionally I worry about my illness
4)	I often worry about my illness
5)	I always worry about my illness
IV.	**Leisure activities**
1)	able to participate in all types of leisure activities without difficulty
2)	able to participate in all types of leisure activities, but with some difficulty
3)	not able to participate in some types of leisure activities
4)	not able to participate in many types of leisure activities
5)	not able to participate in any type of leisure activity
V.	**Pain**
1)	no pain and/or any other complaints
2)	occasionally mild to moderate pain and/or other complaints
3)	often mild to moderate pain and/or other complaints
4)	often moderate to severe pain and/or other complaints
5)	continuous severe pain and/or other complaints
VI.	**Side effects of treatment**
1)	no side effects
2)	occasionally mild to moderate side effects
3)	occasional moderate to severe side effects
4)	often moderate to severe side effects
5)	continuously severe side effects

RESULTS

At baseline the 33 patients who received weekly group physical therapy did not differ significantly from the 26 patients who exercised daily at home regarding age, duration of disease, patient's and physician's global assessment, rating scale and standard gamble utilities, and sociodemographic characteristics (data not shown).
Two patients of the 59 were excluded from analysis. At baseline one patient did not answer the questions seriously, and at followup another patient declined the interview. No interviews were broken off. The mean duration of the interview at baseline was 9.3 minutes (SD, 2.7) for the rating scale and 11.6 minutes (SD, 4.5) for the standard gamble. Seemingly, there was a learning effect. At the 9 month followup the duration decreased to 7.1 minutes (SD, 2.8) for the rating scale and 9.8 minutes (SD, 8.4) for the standard gamble.
Mean values for utilities and means for other selected outcomes at baseline and changes after followup for the 57 patients with AS are shown in Table 3. Improvements in patients' utilities were statistically insignificant by the rating scale (Wilcoxon signed rank test: $p=0.28$) but significant by standard gamble method (Wilcoxon signed rank test: $p=0.04$). On average rating scale utilities improved for those who had weekly group physical therapy [mean:+4 (SD, 11)] but deteriorated [mean:-3 (SD, 9)] in patients who only exercised at home. This difference was statistically different between both groups (Mann-Whitney U test: $p=0.02$). The mean standard gamble utility improved equally (0.03) in both intervention groups.

Reliability of the MUMQ
At followup the one week test-retest reliability was assessed in a convenience sample of 14 stable patients with AS: those patients willing to have a repeated utility assessment after 1 week. For the rating scale the intraclass correlation coefficient was 0.70 for the marker state of mild disease, and 0.95 for patient's own health status. For the standard gamble it was 0.77 for the marker state of mild disease, 0.79 for the marker state of severe disease, and 0.79 for patient's own health status.
Test-retest reliability was *also* assessed for all 57 patients with AS by their valuation of the marker states at followup compared to baseline values. By definition the utility of the marker states should not change over time.[20] Both on the rating scale and on the standard gamble the median of the mild marker state did not change significantly after 9 months (Table 4).

Table 4. Reliability of marker states of disease (n=57)

Rating scale utility (0-100)		
Mild marker		
median change	(range)	5 (-25 - 35)
Standard gamble utility (0-1)		
Mild marker		
median change	(range)	0 (-0.3 - 0.5)
Severe marker		
median change	(range)	0 (-0.9 - 0.95)

Within group comparison: Wilcoxon signed rank test: NS.

Table 3. Utilities and other selected outcomes at baseline and the change (followup-baseline) after 9 months followup for 57 patients with AS

	Baseline Mean	SD	Change Mean	SD
Utilities:				
Patient's valuation of own state of health:				
Rating scale utility	73	15	1	11
Standard gamble utility	0.84	0.19	0.03	0.12
Patient's valuation of marker states of disease:				
Rating Scale mild marker	74	10	3	11
Standard Gamble mild marker	0.87	0.16	0.00	0.15
Standard Gamble severe marker	0.38	0.34	0.07	0.3
Outcome measures:				
AIMS dimensions:				
Mobility	0.25	0.7	0.0	0.24
Physical activity	3.7	2.2	-0.15	2.0
Dexterity	0.8	1.3	0.07	0.87
Social role	0.3	0.7	-0.10	0.45
Social activities	4.4	1.6	-0.38	1.09
Activities of daily living	0.1	0.4	-0.02	0.3
Pain	4.0	2.3	-0.29	1.2
Depression	1.9	1.4	-0.29	1.0
Anxiety	3.1	1.7	-0.3	1.1
Health perception	2.6	1.8	-0.54	1.46
Arthritis impact*	7.3	1.7	0.06	1.61
SIP	4.1	5.8	-0.86	4.4
HAQ-S	0.35	0.3	-0.03	0.12
Functional index	8.4	4.4	-1.19	3.96
Articular index	3.8	3.7	-0.02	3.22
Enthesis index	1.6	4.2	-0.6	2.5
Chest expansion	3.6	1.9	0.55	1.1
Flexion/extension	5.1	2.8	0.64	0.89
Cervical rotation	85	35	19.6	18.5
Physical fitness	152	47	-139	41.3
Pain (VAS)	37.5	26.5	-3.3	22.6
Stiffness (VAS)	36.3	23.4	-1.02	18.4
Physician's global	2.7	0.8	--.--	--.--
Change in patient's global health (VAS)	--.--	--.--	13.8	14.1

* Arthritis impact or patient's global assessment (0=worst and 10=best)

Rating scale utilities versus standard gamble utilities
The baseline utilities obtained with the rating scale did not correlate significantly ($r=0.16$, $p>0.05$) with the utilities obtained by standard gamble at baseline. Neither did the change scores obtained by both techniques ($r=0.11$, $p>0.05$) (Table 5).

Correlations between utilities and patient characteristics
Utilities were correlated with age, duration of disease, sex, marital status and education. Of the demographic characteristics only education correlated significantly with both rating scale and standard gamble, indicating that a higher education was associated with higher utility values. Marital status correlated significantly with utilities obtained with standard gamble method, implying that higher utility values were associated with being married, rather than being unmarried or living together (Table 5).

Construct validity of the MUMQ
- *Spearman correlation.* Baseline rating scale values correlated better with scores on AIMS, SIP, HAQ-S, functional, articular and enthesis indices, pain and stiffness, and spinal mobility than standard gamble utilities (Table 5). This table also indicates which correlations occurred in the "wrong" (paradoxical) direction, although this is in fact a value statement. For example, one would not *a priori* expect a decrease in spinal mobility to be associated with higher utility values. The rating scale correlated significantly with 7 (of 11) dimensions of the AIMS, with the SIP and HAQ-S, with the functional and articular indices, with thoracolumbar flexion/extension and cervical rotation, and with pain and stiffness. The standard gamble correlated significantly with 4 (of 11) dimensions of the AIMS, and with spinal mobility: chest expansion and thoracolumbar flexion/extension (Table 5). Statistically significant but paradoxical correlations were found between standard gamble and 2 dimensions of AIMS (mobility and social activities)(Table 5), suggesting that less mobility or reduced social activities are associated with higher utilities. In interpreting the results, it should be noted that the significance level has not been adjusted for the number of comparisons made.
- *Multiple regression analysis.* For rating scale utilities only patient's global assessment (on the arthritis impact scale of the AIMS)(59%), pain (VAS)(7%), and the functional index (5%) contributed significantly in explaining variance. Together these 3 variables accounted for 71% of total variance in rating scale utility scores (Table 6). For the standard gamble utility, only patient's global assessment (on the arthritis impact scale of the AIMS) (11%) and the SIP (7%) explained significantly total variance. However, the latter occurred in a paradoxical direction (Table 6).

Table 5. Spearman correlation coefficients between both baseline utilities and selected health status outcomes at baseline and after followup

	Rating Scale Utility		Standard Gamble Utility	
	Baseline	Change	Baseline	Change
Age	-0.003	--.--	0.11	--.--
Duration of disease	-0.08	--.--	-0.01	--.--
Sex	0.13	--.--	0.15	--.--
Marital status	-0.05	--.--	-0.34***	--.--
Education	0.31**	--.--	0.22*	--.--
Rating Scale utility	--.--	--.--	0.16	0.11
AIMS dimensions:				
Mobility	-0.26*	-0.07	0.25*@	0.16@
Physical activity	-0.41***	-0.10	-0.12	0.07@
Dexterity	-0.16	0.16@	-0.003	-0.15
Social role	-0.15	0.002@	-0.06	-0.15
Social activities	-0.10	0.21@	0.25*@	-0.01
Activities of daily living	-0.08	0.26*@	-0.08	-0.16
Pain	-0.53***	-0.36***	-0.23*	0.09@
Depression	-0.25*	-0.02	-0.28*	-0.02
Anxiety	-0.38***	-0.11	-0.02	-0.04
Health perception	-0.41***	-0.24*	-0.05	-0.09
Arthritis impact or patient's global	0.68***	0.22*	0.09	-0.00
SIP	-0.33**	-0.02	0.17@	0.21
HAQ-S	-0.41***	0.26*@	-0.04	0.08@
Functional index	-0.49***	-0.29*	-0.12	0.15@
Articular index	-0.44***	-0.21	-0.03	-0.06
Enthesis index	-0.09	0.05@	0.09@	0.15@
Chest expansion	0.20	0.14	0.24*	-0.04@
Flexion/extension	0.26*	0.19	0.36***	0.06
Cervical rotation	0.29*	0.26*	0.18	-0.03@
Physical fitness	0.06	0.18	-0.12@	0.03
Pain (VAS)	-0.66***	-0.38***	-0.21	-0.03
Stiffness (VAS)	-0.31**	-0.15	-0.21	-0.10
Physician's global	-0.18	--.--	-0.17	--.--
Change in patient's global health (VAS)	--.--	0.39***	--.--	-0.10@

* $p < 0.05$ ** $p < 0.01$
*** $p < 0.005$ @ paradoxical direction

Sensitivity to change of MUMQ

• *Spearman correlation.* Changes in rating scale utilities correlated significantly with changes in 4 dimensions of the AIMS [activities of daily living (paradoxical direction), pain, health perception, and arthritis impact], and with changes in HAQ-S (paradoxical direction), functional index, cervical rotation, pain (VAS), and patient's global assessment of change (VAS) (Table 5). The changes in standard gamble utilities did not correlate significantly with any of the other selected health status outcomes (Table 5). Again, p values have not been corrected.

Table 6. Stepwise forward regression analyses with utilities as dependent variable and patient characteristics and other baseline assessments as independent variable

Step	Variable Entered	Partial R^2	F	p value
Dependent variable: Rating scale utility				
1	Patient's global*	0.59	59.67	0.0001
2	Pain VAS	0.07	8.67	0.005
3	Functional index	0.05	6.70	0.01
Dependent variable: Standard gamble utility				
1	Patient's global*	0.11	6.03	0.02
2	SIP	0.07	4.19	0.04**

* arthritis impact dimension of the AIMS ** paradoxical direction

• *Multiple regression analysis.* Changes in rating scale utilities as dependent variable and changes in selected outcomes and the intervention as independent variables showed that 10% of total variance was explained by patient's global assessment of change as indicated on a VAS (Table 7). Changes in standard gamble utilities could not be explained significantly by changes in any of the other selected health status outcomes.

Table 7. Stepwise forward regression analysis with change in rating scale utility as dependent variable and changes in other assessments as independent variables

Step	Variable Entered	Partial R^2	F	p value
Dependent variable: Change in rating scale utility				
1	Change in patient's global assessment	0.10	4.6	0.04

DISCUSSION

Construct validity evaluation showed that rating scale utility values correlated better with health status measures (AIMS, SIP, HAQ-S, functional index, pain and stiffness), and disease activity measures (articular index, enthesis index, and spinal mobility) than standard gamble utilities. Moreover, multiple regression analysis indicated that patient's global assessment explained 59% of total variance of the rating scale values compared to only 11% of standard gamble utility values. These results suggest that standard gamble utilities reflect largely different (as yet unknown) aspects of the health status than rating scale utilities, or have indeed low construct validity.

An evaluation of sensitivity to change showed that with regression techniques changes in rating scale utilities were explained to a higher degree than changes in standard gamble utilities. The question remains whether standard gamble utilities are so generic that they can not be explained by a still limited set of disease outcome variables. That is, one might question whether in the field of utility assessment our concept of construct and discriminant validity is valid. Perhaps standard gamble utilities can only be explained by highly individualized patient preference measures such as for instance a McMaster Toronto Arthritis Rheumatism Patient Preference Disability questionnaire (MACTAR)[21] or the Problem Elicitation Technique (PET) questionnaire.[22] Both questionnaires assess patient priorities, allowing each patient to rank items, i.e., activities affected by arthritis which are of high importance to them. When such measures could also not explain variance in standard gamble utility scores, then in our opinion, the validity of standard gamble utilities may be low and perhaps largely reflect measurement error.

Our findings support the view that rating scale utilities more closely resemble global assessment than a true utility instrument. Indeed, by the rating scale health states are valued under certainty as opposed to the risk assumption inherent to the standard gamble method. Therefore, the numerical quantities obtained by the rating scale are not really von Neumann-Morgenstern utilities. Torrance has suggested that these numerical quantities should in fact properly be referred to as *values* (as opposed to utilities) or *approximations* of von Neumann-Morgenstern utilities.[20] In contrast, the standard gamble method by definition directly measures von Neumann-Morgenstern utilities.[20]

In conclusion, both the rating scale and standard gamble methods of obtaining utilities appeared feasible and reliable. Construct validity testing supports the view that rating scale values address different aspects of health status than do standard gamble utilities. Multiple regression analysis indicated that the rating scale values are strongly related to global assessment results. Multiple

regression analyses of change scores showed higher discriminant validity for the rating scale values than for values obtained by standard gamble technique. Clearly, utility measurement is sensitive to the method chosen to elicit patient well being. This has important implications for decision making and health policy. In our view more validity testing and standardization are needed before utility measurement can be applied fruitfully on a large scale in clinical practice or in health service research.

Acknowledgements
We thank all patients for their cooperation in this study and Cootje Braakman, Ben Dijkmans, Xandra Gielen, Carla Jonker, Annemiek Mackaay, Jacqueline Mertens, Daniël Moolenburgh, Hubert Schouten and Jan Thomassen for their contributions in various stages of the study.

REFERENCES

1 Feeny D, Labelle R, Torrance GW: Integrating economic evaluations and quality of life assessments. In:Spilker B, ed. Quality of life assessments in clinical trials. New York: Raven Press, 1990.

2 Bombardier C, Ware J, Russell J, Larson M, Chalmers A, Read, L: Auranofin therapy and quality of life in patients with rheumatoid arthritis. Am J Med 1986;81:565-78.

3 O'Brien BJ, Elswood J, Calin A: Willingness to accept risk in the treatment of rheumatic disease. J Epidemiol Community Health 1990;44:249-52.

4 Hidding A, van der Linden S, Boers M, et al: Is group physical therapy superior to individualized therapy in Ankylosing Spondylitis? A randomized controlled trial. Arthritis Care Res 1993;6:117-25.

5 Linden van der S, Valkenburg HA, Cats A: Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.

6 Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991;(suppl)12:118-28.

7 Bakker CH, Rutten M, Doorslaer van E, Bennett K, Linden van der S: Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. J Rheumatol 1994;21:269-74.

8 Torrance GW: Social preferences for health states: an empirical evaluation of three measurement techniques. Socioecon Planning Sci 1976;10:129-36.

9 von Neumann J, Morgenstern O: Theory of games and economic behavior, Princeton: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).

10 Torrance GW: Measurement of health-state utilities for economic appraisal: A review. J Health Econ 1986;5:1-30.

11 Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.

12 Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ. Evaluation of the Dutch Arthritis Impact *Measurement Scales* (Dutch-AIMS) in patients with rheumatoid arthritis. Br J Rheumatol 1989; 28:487-91.

13 Daltroy LH, Larson MG, Roberts NW, Liang MH: A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. J Rheumatol 1990;17:946-50.

14 Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B: Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988;15:302-7.

15 Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC: Studies with an enthesis index as a method of clinical assessment in ankylosing spondylitis. Ann Rheum Dis 1987;46:197-202.

16 Miller MH, Lee P, Smythe HA, Goldsmith H: Measurement of spinal mobility in sagittal plane: New skin contraction technique compared with established methods. J Rheumatol 1984;11:507-11.

17 Tugwell P, Bombardier C: A methodological framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982;9:758-62.

18 Fleiss JL: The design and analysis of clinical experiments. New York: Wiley, 1986.

19 Pocock SJ: Clinical trials. A practical approach. New York: Wiley, 1983.

20 Torrance GW, Feeny D: Utilities and quality-adjusted life years. Intl J of Technol Assess Health Care 1989;5: 559-75.

21 Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B: The MACTAR patient preference disability questionnaire: an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987;14:446-51.

22 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.

6

PATIENT UTILITIES IN FIBROMYALGIA AND THE ASSOCIATION WITH OTHER OUTCOME MEASURES

Carla Bakker, Maureen Rutten[1], Marijke van Santen-Hoeufft, Paulien Bolwijn, Eddy van Doorslaer[2], Kathryn Bennett[3], Sjef van der Linden

Department of Internal Medicine, Division of Rheumatology, and the Department of Health Economics[1], University of Limburg, Maastricht, Erasmus University Rotterdam[2], The Netherlands; Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada[3]

Submitted for publication

PATIENT UTILITIES IN FIBROMYALGIA AND THE ASSOCIATION WITH OTHER OUTCOME MEASURES

ABSTRACT

Objective. To compare in patients with fibromyalgia utilities derived by rating scale and standard gamble methods; to gain insight into construct validity by relating utility values to other outcome measures; to assess the sensitivity to change of utilities.

Methods. A total of 73 patients with fibromyalgia were randomized into one of 3 groups: low-impact fitness training, biofeedback, or controls. At baseline and after 6 months the Maastricht Utility Measurement Questionnaire (MUMQ) was applied. By means of both the rating scale and standard gamble method patients were asked to value their own health status. Construct validity of patient utility measurement was evaluated by Spearman correlation and multiple regression of baseline values with pain, stiffness, patient's global assessment, Sickness Impact Profile (SIP), modified Health Assessment Questionnaire (mHAQ), and Arthritis Impact Measurement Scale (AIMS). Sensitivity to change was assessed against changes in these outcomes.

Results. Rating scale utilities correlated significantly ($p < 0.05$) with patient's global assessment ($r_s = 0.53$), pain ($r_s = -0.47$), SIP ($r_s = -0.43$) and with 9 of 11 dimensions of the AIMS (r_s ranging from 0.23 to 0.62). Standard gamble utilities correlated significantly with mobility, pain and arthritis impact of the AIMS scale (r_s from 0.22 to 0.36) and with pain (VAS) ($r_s = -0.24$) and patient's global assessment ($r_s = 0.32$). Multiple regression analysis showed, that patient's global assessment explained 41% (rating scale) and 10% (standard gamble) of total variance in baseline utilities. Also 16% of the variance in *change* in rating scale utility values was explained by changes in patient's global assessment. In contrast, variance of changes in standard gamble utility values was not explained significantly by changes in other disease outcomes.

Conclusion. Rating scale utilities correlated more strongly with disease outcome measures than standard gamble utilities. Also, construct validity for the rating scale was better than for the standard gamble. In fibromyalgia utility measurement is sensitive to the method chosen to elicit patient priorities.

INTRODUCTION

Fibromyalgia is a common rheumatic condition. Presenting symptoms are pain, stiffness and fatigue.[1] It has been suggested that patients with fibromyalgia may benefit from cardiovascular fitness training[2], whereas another report indicated that myobiofeedback training was successful.[3] In these studies conventional disease oriented endpoints such as pain, stiffness, number of tender points, sleep disturbance, fatigue, psychological and global assessments, were measured.
Recently, utility measurement has been introduced in the evaluation of interventions for arthritis patients.[4] Utility measures of health related quality of life are generic measures of the value or preference that patients attach to their overall health status: i.e., patients have to integrate all positive and negative effects of their disease and its treatment into one single value. In contrast, in disease specific instruments the benefits and risks are measured separately. Therefore, one has no information on the relative weights patients assign to therapeutic improvements and disadvantages such as side effects.[5]
In a randomized controlled trial we evaluated the therapeutic effect of low-impact fitness training and biofeedback training on patients with fibromyalgia.[6] During this trial we elicited also utility values. Here we report on aspects of validity of the utility measurement. We address in particular reliability, construct validity and sensitivity to change relative to improvements in other outcomes.

PATIENTS AND METHODS

Study population
Patients with fibromyalgia from the outpatient department of rheumatology of the Maastricht University Hospital, who had been referred between January 1988 and December 1989, were asked to participate in the study. Altogether 103 of 174 (59%) patients gave informed consent of whom 86 met the eligibility criteria (female sex, age 18-60 years, criteria of Wolfe et al.[1]). Patients were excluded if they had a high depression on 6 scales of the symptom check-list (SCL-90).[7,8] For practical reasons the study was restricted to female patients only.

Study design
After baseline assessment patients were randomized into 3 groups. One group received low-impact fitness training, the second group had biofeedback training and the last group were controls. Patients in the fitness group performed supervised aerobic and stretching exercises for 60 minutes, twice weekly, during 6 months. Patients in the biofeedback group individually received 20 minutes relaxation training twice a week, during 2 months.[9] After completing the supervised biofeedback training, patients were encouraged to continue relaxation exercises at home, twice a day for at least 4 more months. Controls received no specific therapy. All patients were allowed to continue the treatment they already received before the study.

Utility measures
To elicit utility values the Maastricht Utility Measurement Questionnaire (MUMQ)[10] was administered at baseline and after 6 month followup by 2 trained interviewers. On both occasions each patient was assessed by the same interviewer who was blinded as to the intervention. The MUMQ, a Dutch translation and adapted version of the McMaster Utility Measurement Questionnaire[11], has been described in detail elsewhere.[10] Briefly, health is

defined by 6 dimensions: activities of daily living, self-care functions, discomfort, leisure activities, pain, and side effects of treatment. Each dimension consists of 5 levels of severity: level 1 reflecting the best situation and level 5 the worst. Marker states were created as follows: perfect health was the combination of the first levels of all 6 dimensions, a severe case of fibromyalgia (marker state of severe disease) was described as the combination of all fifth levels of the 6 dimensions. A mild fibromyalgia marker state was likewise composed.[10]
In the interview the patients were asked to define their own health status by indicating their actual personal levels for each dimension. Then patients were asked to *value* the provided marker states of disease and their own health status, using both the rating scale and standard gamble method. The rating scale is a numerical scale and looks like a thermometer with 'perfect health' equal to 100 at the top and the 'marker state of severe disease' equal to 0 at the bottom. The standard gamble is performed with a probability wheel as a prop.[12] The standard gamble is directly based on the Von Neumann-Morgenstern utility theory and is the original method of measuring utilities.[13] In the standard gamble method health states are valued under the assumption of *risk*, as opposed to the rating scale method where risk is not included in the measurement process.
Usually, patients in the standard gamble utility measurements are asked to value their own state of health in comparison to perfect health (valued as '1') and death (valued as '0') (Figure 1a). However, in rheumatic diseases direct confrontation with the risk of dying may be inappropriate. Therefore, a 2-step utility assessment was performed: patients were first asked to value their own health status in comparison to perfect health and the marker state of *severe* disease (Figure 1b). Next, they were asked to value the marker state of severe disease in comparison to perfect health and death (Figure 1c). The first step provides a standard gamble *score*, which has to be converted into a *utility* value for the patient's **own** health status, using the utility of the severe marker state.[14] Negative utility values for the severe marker state, indicating that this state was considered worse than death, were recoded to zero in the analysis.
Six other health status outcome measures were applied at baseline and followup: global health (on a 0-10 numerical rating scale with 0 equal to 'very bad health' and 10 equal to 'very good health'); a standardized Dutch version of the Sickness Impact Profile (SIP) questionnaire,[15,16]; Dutch Arthritis Impact Measurement Scale (Dutch-AIMS)[17]; pain (on a 10 centimeter visual analogue scale (VAS) with 0 equal to 'no pain' and 10 equal to 'most severe pain imaginable'); duration of morning stiffness (minutes); modified Health Assessment Questionnaire (mHAQ).[18] Note that global health was measured in 2 ways: on the arthritis impact dimension of the AIMS (global health (AIMS)) and on a 0-10 numerical rating scale (global health (NRS)).

Analysis and statistical methods
Reliability was tested in all patients by assessing the stability of marker states after 6 months.
Construct validity of the MUMQ was tested 1-sidedly by Spearman correlation coefficients between *baseline* utility values and *baseline* scores for global health, SIP, AIMS, mHAQ, pain and stiffness. Construct validity indicates whether method results do agree with expected results based on *a priori* assumptions of the investigator.[19] In addition, multiple regression analyses (stepwise forward) were performed for these 6 clinical measures and age, disease duration, marital status and education as independent variables and the rating scale or standard gamble utilities as dependent variables. Independent variables with skewed distributions were analyzed as ln(var+1), the natural logarithm of one plus the variable.[20,21]

Figure 1a. Standard Gamble: Value your own state of health in comparison to perfect health and death

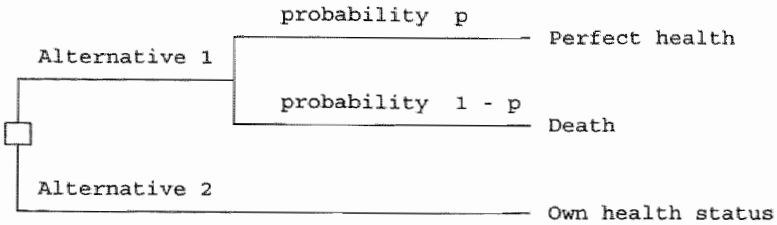


Figure 1b. Two-step Standard Gamble: first step: Value your own state of health in comparison to perfect health and the marker state of severe disease

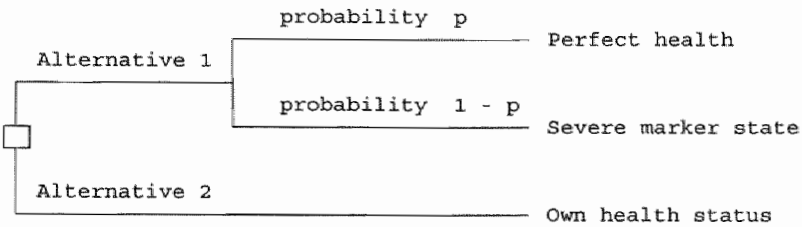
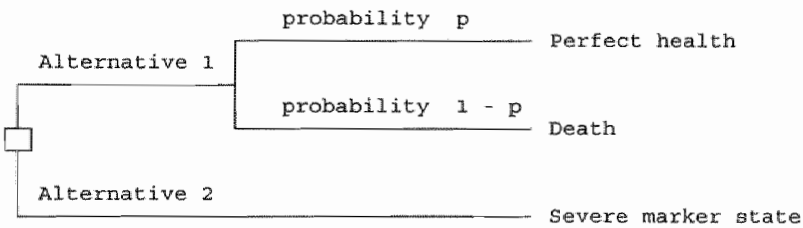


Figure 1c. Two-step Standard Gamble: second step: Value the marker state of severe disease in comparison to perfect health and death



Discriminant validity or sensitivity to change indicates whether a measure can detect important clinical changes in health status over time.[19] Discriminant validity was tested in 2 ways. First, we calculated Spearman correlation coefficients (1-sided testing) between the changes in utility values and the changes in other outcome measures. Secondly, we performed multiple regression analyses (stepwise forward) with changes in rating scale or standard gamble utilities as dependent variable and treatment and changes in the other health status outcomes as independent variables. Both dependent and independent variables were transformed by ln(var+constant), the

natural logarithm of a constant just below the minimum of (change in) the variable plus (change in) the variable. Within each treatment group mean changes in utility were tested by Wilcoxon signed rank test.

RESULTS

At baseline all 86 eligible patients with fibromyalgia completed the MUMQ. One patient withdrew before randomization and 12 patients dropped out during the 6 month followup (6 in the fitness group, 5 in the biofeedback group; 1 from the control group) for the following reasons: illness of husband (2), too busy job (2), hospitalization (1), no further interest (4), trial too stressful (2), biofeedback makes no sense (1). The remaining 73 patients (fitness group 29; biofeedback group 26; controls 18) are included in this report. Demographic and clinical characteristics are shown in Table 1. Patients in the fitness group were significantly older than the controls (44.9 years compared to 40.1 years) ($p=0.05$). There were no other statistically significant differences in baseline characteristics and utilities between the 3 intervention groups. Four patients gave inconsistent answers at baseline. These answers were excluded from analysis.[10] No interviews were broken off. The mean duration of the utility measurement at baseline was 10.8 minutes (SD, 2.9) for the rating scale and 12.5 minutes (SD, 3.8) for the standard gamble; at the 6 month followup it decreased to 9.4 minutes (SD, 7.3) for the rating scale and 11.5 minutes (SD, 4.5) for the standard gamble method.

Table 1. Baseline characteristics for the 73 patients with fibromyalgia

Age (years)		
mean (SD)		43.3 (8.3)
Duration of complaints (years)		
mean (SD)		11.8 (9.8)
% Married		81
% Employed		27
Educational level	% high*	37
	% low**	63

* including secondary vocational training and university
** including lower vocational training

Mean values for utilities and other outcomes are shown in Table 2. Patient's utilities improved significantly by rating scale only (Wilcoxon signed rank test: $p=0.008$). These utilities showed a significant improvement over time for those who had low-impact fitness training (mean improvement 11; $p=0.007$) but were insignificant among patients who had biofeedback training (mean improvement 3; $p=0.3$) and stayed about the same in controls (mean improvement 0.1; $p=0.99$). Differences between the three groups did not differ significantly (Kruskal-Wallis test: $p=0.32$). Standard gamble utilities however, did not change significantly in either the fitness group (mean: +0.06; $p=0.2$), among controls (mean: +0.01; $p=0.9$) or in the biofeedback group (mean: -0.01; $p=0.8$).

Reliability

Test-retest reliability was assessed by the patient's valuation of the *marker states* at the 6 month followup visit compared to baseline values. Utilities of marker states should not change over time.[22] However, on the rating scale the mean of the marker state values for mild disease increased significantly (mean change: 4.0; Wilcoxon signed rank test: p=0.03). This was not the case in the standard gamble (mean change: 0.02; p=0.6), but the mean standard gamble utility for the marker state of severe disease deteriorated significantly (mean change: -0.11; p=0.01) (Table 2). It should be noted that on the rating scale the bottom endpoint is the severe marker state and therefore, the test-retest reliability of this state could only be tested by the standard gamble method.

Table 2. Utilities and other outcomes at baseline and changes at followup among 73 patients

	Baseline		Change*	
	Mean	SD	Mean	SD
Utilities:				
Patient's valuation of **own state of health**:				
Rating Scale utility	55	18	5.5	20.3
Standard Gamble utility	0.81	0.20	0.02	0.25
Patient's valuation of **marker states of disease**:				
Rating Scale mild marker	73	13	4.0	13.4
Standard Gamble mild marker	0.83	0.21	0.02	0.25
Standard Gamble severe marker	0.42	0.38	-0.11	0.35
Outcome measures:				
AIMS dimensions:				
Mobility	0.5	1.2	0.22	1.42
Physical activity	5.5	2.0	0.03	2.33
Dexterity	3.2	2.7	-0.08	2.56
Social role	0.7	0.8	0.23	0.65
Social activities	4.3	1.7	-0.51	1.16
Activities of daily living	0.3	0.9	0.05	0.72
Pain	7.1	1.5	-0.30	1.52
Depression	3.7	1.6	-0.29	1.55
Anxiety	5.2	1.7	-0.37	1.45
Health perception	4.1	1.9	-0.26	1.60
Arthritis impact or global health	5.4	2.1	0.15	2.38
SIP	14	8.6	-1.36	6.55
mHAQ	0.47	0.37	0.07	0.37
Pain (VAS)**	5.9	1.9	-2.75	17.3
Stiffness	65	57	12.3	46.9
Patient's global health (NRS)***	5.4	1.4	0.85	1.76

* (followup - baseline) ** visual analogue scale *** numerical rating scale

Comparison of rating scale and standard gamble utilities
Utilities did not correlate significantly with age, duration of disease, marital status or education. Utilities obtained by rating scale did not correlate significantly ($r=0.14$, $p>0.05$) with utilities obtained by standard gamble. The same was found for change scores ($r=0.19$, $p>0.05$) (Table 3).

Construct validity of utility assessment
• *Spearman correlation.* Rating scale scores correlated better with scores on AIMS, SIP, mHAQ, pain, stiffness, and global health than standard gamble scores (Table 3). Rating scale utilities correlated significantly with the arthritis impact, physical activity, social role, pain, depression and health perception dimensions of the AIMS and also with SIP, mHAQ, global health (NRS) and pain (VAS) (Table 3). Standard gamble utility values correlated significantly with the arthritis impact, mobility and pain dimensions of the AIMS and with the global health (NRS) and pain (VAS) scales only (Table 3). In interpreting the results, it should be noted that the significance level has not been adjusted for the number of comparisons made.

Table 3. Spearman correlation coefficients between baseline utilities and health status outcomes at baseline and after followup for 73 patients with fibromyalgia

	Rating Scale utility		Standard Gamble utility	
	Baseline	**Change**	**Baseline**	**Change**
Age	0.08	--.--	0.20	--.--
Duration of disease	-0.01	--.--	-0.06	--.--
Marital status	0.01	--.--	-0.01	--.--
Education	0.21	--.--	0.001	--.--
Rating Scale	--.--	--.--	0.14	0.19
AIMS dimensions:				
Mobility	-0.39***	-0.17	-0.22*	0.03@
Physical activity	-0.52***	-0.18	-0.16	-0.13
Dexterity	-0.17	0.06@	-0.21	0.07@
Social role	-0.28*	0.20@	-0.15	-0.04
Social activities	-0.23*	-0.003	-0.03	-0.07
Activities daily living	-0.26*	-0.19	-0.08	0.09@
Pain	-0.38***	-0.33**	-0.22*	-0.09
Depression	-0.30*	-0.27*	-0.20	-0.24*
Anxiety	-0.14	-0.29*	-0.16	-0.09
Health perception	-0.29*	-0.07	-0.19	0.07@
Arthritis impact or global health	0.62***	0.41***	0.36***	0.23*
SIP	-0.43***	-0.23*	-0.08	0.06@
mHAQ	-0.28*	0.09@	-0.16	0.003@
Pain (VAS)$	-0.47***	-0.25*	-0.24*	0.01@
Stiffness	-0.19	-0.15	-0.09	-0.05
Patient's global health (NRS)#	0.53***	0.41***	0.32**	0.24*

* $p<0.05$ ** $p<0.01$ *** $p<0.001$
@ paradoxical direction $ visual analogue scale # numerical rating scale

• *Multiple regression analysis*. A multiple regression analysis was performed to determine which set of variables could best predict the rating scale and standard gamble utilities. For rating scale utilities, patient's global health (AIMS) explained 41% of the variance and the physical activity dimension (AIMS) another 11%. Variance in standard gamble utilities was explained significantly only by patient's global health (AIMS) for 10% (Table 4).

Table 4. Stepwise forward regression analyses with rating scale or standard gamble utilities as dependent variable and patient characteristics and other baseline assessments as independent variables

Step	Variable entered	Partial R^2	F	p value
Dependent variable: Rating scale utility				
1	Global health (AIMS)	0.41	33.44	0.0001
2	Physical activity (AIMS)	0.11	11.11	0.002
Dependent variable: Standard Gamble utility				
1	Global health (AIMS)	0.10	5.88	0.02

Sensitivity to change of utility assessment

• *Spearman correlation*. Changes in rating scale utilities correlated significantly with changes in 4 dimensions of the AIMS (pain, depression, anxiety, arthritis impact), and with changes in *SIP*, pain (VAS), and patient's global health (NRS) (Table 3). Note that this table also indicates which correlations occurred in the "wrong" (paradoxical or unexpected) direction. For example, one would not a priori expect a decrease in social functioning to be associated with higher utility values. Changes in standard gamble utilities correlated significantly with changes in depression and arthritis impact dimensions of the AIMS, and with changes in global health (NRS). Again, p values have not been corrected for the number of tests performed.

• *Multiple regression analysis*. Multiple regression analysis with changes in rating scale utilities as dependent variable and changes in the other outcomes and treatment (group) as independent variables showed that 16% of total variance could be explained by changes in patient's global health (AIMS)(Table 5). Changes in standard gamble utilities could not be expiained significantly by changes in any of these variables.

Table 5. Stepwise forward regression analyses with changes in rating scale or standard gamble utilities as dependent variables and changes in other assessments as independent variables

Step	Variable entered	Partial R^2	F	p value
Dependent variable: Change in Rating scale utility				
1	Global health (AIMS)	0.16	9.24	0.004
Dependent variable: Change in Standard Gamble utility				
1	mHAQ	0.07	3.61	0.06*
2	Physical activity (AIMS)	0.06	3.23	0.08
3	Social activity (AIMS)	0.06	3.16	0.08

* paradoxical direction

DISCUSSION

We evaluated reliability, construct validity, and sensitivity to change of utility measurement by rating scale and standard gamble method. These aspects of utility measurement differed considerably between both methods. Therefore patient's utilities elicited by rating scale and standard gamble are not interchangeable.

The test-retest *reliability* of the MUMQ was assessed by the utilities for the marker states of disease that of course should not change over time.[22] The utilities of these marker states were in fact not stable (Table 2). Therefore, either the method itself has poor reliability, *or* the patient's perception and valuation of the marker states indeed changes in the course of 6 months. Note that both the valuation of the patient's own health status and the valuation of the mild marker state changed in the same direction, i.e., they were at a higher mean level after 6 months (Table 2). Moreover, as the patient's utility improved, the distance between her own health status and the marker state of severe disease became larger. Patients emphasized this by valuing the marker state of severe disease lower at followup compared to baseline (Table 2). Therefore, possibly a change in the patient's perception of her own health status induces valuations for the reference (marker) states to change too. We suggest that this might be related to a patient's capabilities to adapt to disease, i.e., she might be better able to deal with her disease related limitations, and the perceptions of other patients with the same disease may change too. As marker states are presented as examples of mild and severe fibromyalgia, the valuation of these reference states may change accordingly. Future research should clarify this issue.

Construct validity of utilities obtained by rating scale was supported by significant correlations with measures such as global health, pain, SIP, AIMS, and mHAQ. Standard gamble utility values, however, correlated considerably less with these instruments. Patient's global health explained 41% and 10% of total variance of rating scale and standard gamble utilities respectively. This suggests that standard gamble utilities reflect different aspects of health status than rating scale utilities or have indeed considerably lower construct validity. In patients with ankylosing spondylitis construct validity appeared also to be higher for the rating scale than for the standard gamble.[23] Clearly, the 2 techniques are not interchangeable. It should be stressed that the standard gamble method incorporates a risk of getting a dispreferred outcome, whereas risk is not an issue in the rating scale procedure. The standard gamble method therefore, addresses at least elements of uncertainty.

Our findings support the view that utilities obtained by rating scale more closely resemble global assessment. The differences between rating scale and global assessment relate to the endpoints of these scales. Global assessments are measured in many (flexible) ways, i.e., with a variety of different endpoints. In contrast, rating scale utilities are measured in a standardized way with perfect health and (usually) death as endpoints. Therefore, the methodological advantage of standardized rating scale utility measurement over non-standardized global assessment is that utilities provide numerical values which allows patient outcomes of different diseases or resulting from various health care interventions to be compared across patients and diseases.

An evaluation of *sensitivity to change* showed that changes in rating scale utilities could be explained to a higher degree than changes in standard gamble utilities.

In conclusion, reliability of utility measurement by rating scale and standard gamble method assessed by stability of marker states was rather poor in patients with fibromyalgia. Correlations between utilities and other outcomes showed higher construct validity and sensitivity to change for the rating scale than for standard gamble utilities. Regression analysis indicated that rating scale values are strongly related to global assessment results. Rating scale utilities are better standardized than many global assessments, and they can be compared across patients, treatments and diseases. Clearly, utility measurement is sensitive to the method chosen to elicit patient well being. This has important implications for decision making and health policy. In our view, more validity testing and standardization are needed before utility measurement can be applied on a larger scale in clinical practice or in health service research.

Acknowledgments

We thank all patients and Rob de Bie, Annemiek Fransen, Nico Groenman, Paul Kubben, Ton Lenssen, Hubert Schouten, Carlo Theunissen, Vicky Verstappen and Frans Verstappen for their contributions in various stages of the study.

REFERENCES

1 Wolfe F, Smythe HA, Yunus MB, et al: The American College of Rheumatology 1990 Criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. Arthritis Rheum 1990;33:160-72.

2 McCain G, Bell D, Mai F, Halliday P: A controlled study of the effects of a supervised cardiovascular fitness training program on the manifestations of primary fibromyalgia. Arthritis Rheum 1988;31:1135-41.

3 Ferraccioli G, Ghirelli L, Scita F, et al: EMG-Biofeedback training in fibromyalgia syndrome. J Rheumatol 1987;14:820-5.

4 Bombardier C, Ware J, Russell J, Larson M, Chalmers A, Read, L: Auranofin therapy and quality of life in patients with rheumatoid arthritis. Am J Med 1986;81:565-78.

5 Feeny D, Labelle R, Torrance GW: Integrating economic evaluations and quality of life assessments. In: Spilker B, ed. Quality of life assessments in clinical trials. New York: Raven Press, 1990.

6 Santen van-Hoeufft M, Bolwijn P, Kleijnen J, Verstappen F, Bakker C, Hidding A, Houben H, Linden van der Sj: Is low or high impact fitness beneficial in fibromyalgia patients. Results of 2 randomized controlled trials. (in preparation)

7 Arrindell WA, Ettema JHM: Handleiding bij een multidimensionele psychopathologie-indicator. Swets & Zeitlinger B.V., Lisse, 1986.

8 Derogatis LR: SCL-90: Administration, scoring and procedures manual-I for the r(evised) version. Baltimore: Johns Hopkins University School of Medicine, Clinical Psychometrics Research Unit, 1977.

9 Basmajian JV eds. Biofeedback, principles and practice for clinicians. Baltimore: Williams and Wilkins Company, 1979.

10 Bakker CH, Rutten M, Doorslaer van E, Bennett K, Linden van der S: Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. J Rheumatol 1994;21:269-74.

11 Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991;(suppl)12:118-28.

12 Torrance GW: Social preferences for health states: an empirical evaluation of three measurement techniques. Socioecon Planning Sci 1976;10:129-36.

13 von Neumann J, Morgenstern O: Theory of games and economic behavior. Princeton: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).

14 Torrance GW: Measurement of health-state utilities for economic appraisal: a review. J Health Econ 1986;5:1-30.

15 Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.

16 Luttik A, Jacobs H, de Witte L: Een Nederlandse versie van de Sickness Impact Profile. Vakgroep Huisartsgeneeskunde. Rijksuniversiteit Utrecht, 1987 (2nd ed.).

17 Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ: Evaluation of the Dutch Arthritis Impact Measurement Scales (Dutch-AIMS) in patients with rheumatoid arthritis. Br J Rheumatol 1989;28:487-91.

18 Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP: Assessment of patient satisfaction in activities of daily living using a modification of the Stanford health assessment questionnaire. Arthritis Rheum 1983;26:1346-53.

19 Tugwell P, Bombardier C: A methodological framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982;9:758-62.
20 Fleiss JL. The design and analysis of clinical experiments. New York: Wiley, 1986.
21 Pocock SJ. Clinical trials. A practical approach. New York: Wiley, 1983.
22 Torrance GW, Feeny D: Utilities and quality-adjusted life years. Intl J of Technology Assessment in Health Care 1989;5:559-75.
23 Bakker CH, Rutten M, Hidding A, van Doorslaer E, Bennett K, van der Linden S: Patient utilities in ankylosing spondylitis and the association with other outcome measures. J Rheumatol 1994;21:1298-1304.

7

METHODOLOGICAL ISSUES OF PATIENT UTILITY MEASUREMENT: EXPERIENCE FROM TWO CLINICAL TRIALS

Maureen Rutten[1], Carla Bakker[2], Eddy van Doorslaer[3], Sjef van der Linden[2]

Department of Health Economics[1] and Department of Internal Medicine, Division of Rheumatology[2], University of Limburg, Maastricht, Institute for Medical Technology Assessment, Erasmus University Rotterdam[3], The Netherlands

Accepted for publication in Medical Care

METHODOLOGICAL ISSUES OF PATIENT UTILITY MEASUREMENT: EXPERIENCE FROM TWO CLINICAL TRIALS

ABSTRACT

This paper explores various methodological issues of patient utility measurement in 2 randomized controlled clinical trials involving 85 fibromyalgia and 144 ankylosing spondylitis patients. In both trials one baseline and two followup measurements of patients' preferences for their own health state and several hypothetical states were performed using the rating scale and the standard gamble methods.

It was confirmed that standard gamble scores are consistently higher than rating scale scores for both the experienced and hypothetical states. The three-month test-retest reliability for hypothetical states measured by intraclass correlation coefficients ranged from 0.24 to 0.33 for the rating scale and from 0.43 to 0.70 for the standard gamble. Although the reproducibility is not high, the group mean scores are fairly stable over time. Mean standard gamble scores tend to differ depending on the way the measurements are undertaken. Utilities elicited with chained gambles were significantly higher than utilities elicited with basic reference gambles. At the individual level some inconsistent responses occurred. However, more than 70% of these fell within the bounds of the measurement error which ranged from 0.11 to 0.13 on the standard gamble (0-1 scale) and from 8 to 10 on the rating scale (0-100 scale). The large number of negative utilities for the severe hypothetical state, which was used as an anchor point in the chained gambles, and the magnitude of these negative utilities (down to -19) lead us to favor using death as the anchor point in future applications of the standard gamble method.

INTROCUTION

Decisions regarding the allocation of resources to health care interventions should ideally be based on the relative costs and benefits of the alternatives. This requires an assessment of the societal value of the outcomes (i.e., various health states) achieved by these interventions, which can be done by means of utility measurement. When utility measurement is applied in clinical trials, patients are asked to assign a single value to a health state on a scale ranging from 0 (usually death) to 1 (usually perfect health), by balancing the positive treatment effects

against the negative side effects.[1] An utility can be seen as an inclusive, generic quality of life measure, which reflects the net effect of treatment. It is designed to allow a broad comparison of the effects of health care interventions across patient populations. When such comparisons are made in the context of cost-utility analyses, utilities are often used as weights to compute "quality adjusted life years" (QALYs). Years of life are multiplied by utility weights for the health status during those years, thus adjusting these life years for their quality.

Because the impact of a health care intervention on the value of an individual patient's health state is increasingly recognized as an essential component of the evaluation of that intervention's usefulness, utility measurement has become more widely used during the past decade. In general, two different approaches to utility measurement have been used in clinical trials.[2] In the *first approach*, patients are asked a number of questions about their functioning. Their answers are used to rate them on the various quality of life dimensions of a particular utility measurement instrument. Combining these dimensions results in descriptions of patients' overall health states, to which preference values are assigned. These values are obtained in a different population, usually the general population. The majority of utility analyses following this approach have used the health state preference values that were obtained by the original developers of the utility instruments.[3] The most commonly used pre-packaged utility measurement instruments for this approach are Rosser's Disability Distress scale[4], Kaplan's Quality of Wellbeing scale[5], and Torrance's Health Utility Index.[6]

In contrast to the use of pre-packaged systems, the *second approach* to utility measurement is to ask patients to assign a single preference value to their overall quality of life. This self valuation by patients has not been widely undertaken in the past, but appears to be increasingly used.[3] The Auranofin trial in rheumatoid arthritis patients is a well known example of this approach.[7,8] It has yet to be established whether one of these two approaches or any of the available utility instruments is superior.

In the two studies in fibromyalgia and ankylosing spondylitis, described in this paper, we have opted for the second approach. It seems very appealing to us to incorporate patients' preferences in evaluating a therapy, since only they know the true implications of a particular health state from first hand experience, and their preferences reflect the relative desirability of different health states to those who should benefit from the services provided.[9] Moreover, the first approach commonly requires decision rules to classify patients into the dimensions of a utility instrument using the answers they gave on other quality of life questions. Often these decision rules are not easily explained and justified.

When using pre-packaged systems, the underlying utilities may have been elicited by a number of techniques of which the rating scale, standard gamble and time trade-off are the major ones. These same techniques can also be used in the approach we opted for, where utilities are derived directly from patients. A comprehensive description of these techniques is given by Torrance.[10] In our study only the rating scale and the standard gamble were used. The first because it is a very simple technique in which a subject provides his preference values explicitly by placing health states on a scale with clearly defined endpoints (e.g. best imaginable health state and worst imaginable health state). The second because it is the only technique that is well-founded on an economic theory, the Von Neumann and Morgenstern expected utility theory.[11] We did not consider using the time trade-off technique because on the one hand it seems more difficult than the rating scale and on the other hand it is not founded on a particular theory. Furthermore, for practical reasons the number of techniques had to be limited. A fuller description of the rating scale and standard gamble techniques is given in the methods section.

A recent review by Froberg and Kane of measurement issues related to obtaining utilities,

indicated there was a considerable lack of knowledge about the accuracy of utility measurement techniques.[12a-12d] This paper attempts to make a contribution to overcoming parts of this deficiency by focussing on those methodological issues of patient utility measurement that became apparent in our studies. In the methods section we will provide a brief description of the patients and the studies in which the utility measurements were incorporated, and the methodology of utility measurement that was used. The main body of this paper contains four separate sections on methodological issues. In the first section, a comparison is made between rating scale and standard gamble utilities whereas in the second section the reliability of both techniques is compared. The third section concerns the internal consistency of the standard gamble and the fourth section addresses health states valued worse than death. Some issues, such as the observed differences between rating scale and standard gamble methods[13-16], inconsistent responses and the differences between basic reference gambles and chained gambles[17,18] have been addressed before, but usually not in the context of clinical trials. Reliability data are scarcely reported in the existing literature and little has been reported empirically on the occurrence of negative utilities.

METHODS

Patients and Studies

Utility measurements were performed in the context of 2 randomized controlled clinical trials in rheumatic patients. These trials and their results are described in more detail elsewhere.[19,20] Patients were recruited for both studies from outpatient clinics. In the first study, 85 females suffering from *fibromyalgia* (criteria of Wolfe et al.)[21] were randomly divided into a standardized fitness training program (n=35), a biofeedback training program (n=31) and a control group (n=19). Their mean±SD age was 44±8 years, 82% percent was married, 27% employed and 66% had a low educational level. At the 6 months followup 12 patients dropped out, 6 in the fitness group, 5 in the biofeedback group and 1 in the control group, all of them for reasons unrelated to their disease.[20] The baseline characteristics of the drop-outs did not differ significantly from those of the patients who completed the study.

In the second study 144 patients (21% female) suffering from ankylosing spondylitis (modified New York criteria)[22] were randomly assigned to receive either 1) self-administered unsupervised individual physical exercise at home (n=68) or 2) weekly sessions of group physical therapy in addition to the same individual physical exercise at home (n=76). Their mean±SD age was 43±10 years, and 67% was married, 72% employed and 35% had a low educational level. By the 9 months followup 9 patients had withdrawn (8 in the experimental and 1 in the control group), 4 because of the inability to exercise individually, 4 due to other diseases or pregnancy and 1 who had moved. Their baseline characteristics did not differ significantly from those who did not withdraw.[19]

Since the focus of this paper is on methodological issues, it is not our intention to address the effectiveness of the therapies (details on effectiveness can be found in Hidding et al., and Van Santen-Hoeufft et al.). The trial data are only used to illustrate some methodological issues.

Methods of utility measurement

In the fibromyalgia study patients were seen for utility measurement at an outpatient clinic at baseline and at the 3 and 6 months followups whereas in ankylosing spondylitis utility measurements were scheduled at baseline, 3 and 9 months followups. The measurements were done by means of the Maastricht Utility Measurement Questionnaire, a translated and slightly adapted version of the McMaster Health Utility Index.[23,24] This instrument is administered by a trained interviewer and takes about 30 minutes to complete.

All patients were asked to provide utilities for 3 hypothetical reference health states as well as for their own health state. The description of each health state covers 6 dimensions: 1) activities of daily living, 2) self-care functions, 3) anxiety and depression, 4) leisure activities, 5) pain and discomfort and 6) side effects from treatment. Each patient described his or her own health state by ticking off one of the 5 functional levels within each dimension (1 being the best functional level and 5 being the worst functional level). For all health states, duration was specified as "the rest of your life". The reference states describe typical mild, moderate and severe states of patient's illness. They help patients to determine the position of their own health state on the spectrum of possibilities.[25] Since the reference states remain the same throughout the study, they also enable the calculation of the test-retest reliability of the utility measurement instrument at the repeated followups.[25]

As part of the Maastricht Utility Measurement Questionnaire, utilities were elicited using both the rating scale and the standard gamble method. In the first part of the baseline and followup interviews patients rank the reference health states and their own health state by preference on a rating scale with the endpoints of perfect health (100) and the severe reference state (0). They are asked to do this in such a way that the distances between the states represent the differences in their preferences. The baseline interview then continues with standard gamble questions 1, 2 and 3 as shown in table 1. At followup interviews an additional question (question 4) is asked to assess the change which the patient has experienced relative to baseline (see table 1). Before question 4, each patient is explicitly asked whether his health state has improved, deteriorated or remained stable compared to baseline. If a patient indicates an improvement compared to baseline, question 4a is asked. When a patient indicates a deterioration, question 4b is asked. The first 80 ankylosing spondylitis patients coming to the 9 months followup were asked an additional fifth question in which they valued their own current health state using perfect health and death as outcomes of the gamble. At 3 and 9 months followup all ankylosing spondylitis patients who found that the severe reference state was worse than death were asked a sixth question to assess the magnitude of the negative utility for the severe reference state. Throughout this paper we will refer to the question numbers in table 1.

Table 1. Standard gamble questions at baseline (1,2 and 3) and at followup (1,2,3,4,5,6)*

Standard gamble question	Health state being valued	Outcomes of the gamble	
		Best (p)	Worst (1-p)
1	mild reference state	perfect health	severe reference state
2	own health state	perfect health	severe reference state
3	severe reference state	perfect health	death
4a	own health state at followup	perfect health	own health state at baseline
4b	own health state at baseline	perfect health	own health state at followup
5	own health state	perfect health	death
6	death	perfect health	severe reference state

* question 5 was only put to 80 ankylosing spondylitis patients at 9 months followup; question 6 only at 3 and 9 months followup to ankylosing spondylitis patients who found the severe state worse than death.

The standard gamble is sometimes seen as the gold standard for utility measurement, because it is directly based on the axioms of the Von Neumann and Morgenstern expected utility theory.[11] This theory consists of a number of axioms for rational decision-making under risk. One of these axioms specifies the standard gamble approach to utility measurement. In each standard gamble question a patient is offered a choice between certain continued life in the health state being evaluated (hi), and a gamble with chance p to gain the best outcome of the gamble (perfect health) and chance 1-p of attaining the worst outcome of the gamble. The health state being valued must be intermediate between the two outcomes of the gamble in terms of preference. Chance p is systematically varied until the patient is indifferent between continued life in state hi and taking the gamble. In our studies p was varied in steps of 10% (p/1-p: 100/0, 90/10, 10/90, 80/20, 20/80 etc.) When the indifference point has been found, the utilities of the health states (U_{hi}) are calculated using the expected utility equation: $U_{hi} = pU_x + (1-p)U_z$, where p is the indifference probability, U_x is the utility of the best outcome of the gamble and U_z is the utility of the worst outcome of the gamble.**

** The indifference probability is defined as the midpoint of the two probabilities of perfect health between which the preference shifts from the gamble to the sure state. For example, if a patient prefers a gamble with probabilities 90/10 to the sure health state, but prefers the sure health state to a gamble with probabilities 80/20, the indifference probability is 0.85.

Standard gambles with perfect health and death as potential outcomes are called *basic reference gambles*. Since by definition the utility of perfect health is 1 and the utility of death is 0, the utility of the health state being valued in a basic reference gamble is equal to p.1 + (1-p).0 = p, the indifference probability. Generally, the more undesirable the health state being valued, the greater the willingness to take a risk in order to escape that health state, the lower the indifference probability p, and thus the lower the utility for that state. Thus, the standard gamble provides an implicit valuation of a health state relative to the 2 possible outcomes of the gamble.

The worst (or best) outcome of the gamble can be replaced by any other health state as long as it is worse (or better) than the health state being valued.[26] Such gambles are called *chained gambles*, because they have to be chained to a basic reference gamble that assesses the utility of that other health state. In our studies, the severe reference state was substituted for death, in order to avoid using death in a gamble that involved a chronically ill patient's own health state. Including death could upset them. Moreover, in the period covered by our studies, death was unlikely to be a relevant outcome in the rheumatic disease patient groups we studied. When the severe reference state is used as the worst outcome of the gamble (as in standard gamble question 1 and 2) the utility for the health state being valued can be calculated using the same equation as above, where U_z is the utility of the severe reference state, a utility which is measured in basic reference gamble 3. In standard gamble question 4a, U_z is the utility of the patient's own health state at baseline, which was measured during the baseline-interview. In standard gamble 4b, U_z is the utility of the patient's own health state at followup, which is being measured in the followup interview.

In general, when the severe reference state was considered worse than death, a utility of 0 was assigned to that state in order to avoid using negative utilities. When indifference is reached in standard gamble question 6, the magnitude of the negative utility is calculated as -p/(1-p), where p is again the indifference probability.[10] Although calculated, these negative utilities were not used in the chained gambles.

To facilitate the patients' understanding of the standard gamble questions a probability wheel was used as a visual aid.[15] This is an adjustable disk with two different colored sectors that reflect the probabilities of getting the two outcomes of the gamble. The outcomes of the gamble are described on cards that have the same color as the sectors. The size of the sectors is changed according to the change in probability.

Statistical Analysis

Results will be presented using means and associated standard errors. However, since the negative utilities of the severe reference state elicited in standard gamble question 6 included a number of extremely negative values we present the 5% trimmed mean and the 5% trimmed standard deviation for these results. This means that the upper and lower 5% of all observations were excluded when calculating the mean and standard deviation, thus removing the influence of the outlier values that caused the distribution of negative utilities to be heavily skewed to the left.[27] Within-patient analyses by means of paired t-tests were performed 1) to compare rating scale with standard gamble scores, 2) to compare chained gambles with the basic reference gamble, and 3) to test for differences between reference state scores over time. Pearson product moment correlations are reported as measures of association between rating scale and standard gamble scores. A logistic regression analysis was performed to test for differences between patients who did and those who did not gave inconsistent responses. Intraclass correlation coefficients were calculated to examine test-retest reliability.[28] Reliability was further assessed

in terms of the precision of an individual measurement. This precision, expressed as σ_e, is the standard deviation of the measurement error also called the standard error of measurement.[29] It is calculated as the square root of the mean square error (MSE) which is given by an analysis of variance.[28] It can also be calculated as $\sigma \sqrt{(1-R)}$, where σ is the standard deviation of all measurements and R is the test-retest reliability coefficient.[29]

RESULTS

1. Comparing Rating Scale and Standard Gamble Utilities

Although the standard gamble method is sometimes seen as the gold standard, the rating scale method is far more frequently used, probably because it is less time consuming and easier to apply. In our studies Pearson product moment correlations between rating scale and standard gamble scores for various health states were found to range between 0.31 and 0.48 ($p<0.001$). However, highly significant correlations can coexist with systematic differences between the methods. As can be seen in table 2, the mean utilities of the patients' own health states assessed via the standard gamble method were significantly higher than the utilities assessed via the rating scale method. This pattern, which was also found for the mild reference health state, is consistent with earlier findings.[13-16]

Table 2. Mean (SE) rating scale and standard gamble values for the patient's own health state

	N	Rating scale* endpoints: perfect health - severe	Standard gamble outcomes: perfect health - severe	p value**
Fibromyalgia				
baseline	85	0.54 (0.020)	0.67 (0.028)	0.000
3 months	76	0.57 (0.023)	0.76 (0.023)	0.000
6 months	73	0.60 (0.022)	0.76 (0.025)	0.000
Ankylosing spondylitis				
baseline	144	0.72 (0.013)	0.75 (0.018)	0.095
3 months	137	0.74 (0.012)	0.78 (0.016)	0.009
9 months	133	0.75 (0.013)	0.79 (0.015)	0.002

* Rating scale preferences were divided by 100 ** Paired t-test

In ankylosing spondylitis we found a difference of somewhat less than 5% between the methods, and in fibromyalgia we found a difference of more than 10%. Such differences might considerably affect the results of a cost-utility analysis and alter the conclusions drawn. Whether this happens depends on the sensitivity of the decisions to the observed range of variation. Some cost utility ratios may be very robust to the magnitude of the utility, whereas others may even change as a result of only a very small change in utility. The variability of responses among patients is somewhat greater for the standard gamble method than for the rating scale method. Several phenomena might explain the differences between rating scale and standard gamble preferences. Three of them are discussed below.

Response Spreading

The first is called "response spreading" on the rating scale. This means that patients tend to distribute the health states over the entire scale, even if the true values were bunched at one end.[14,30]. The mean baseline rating scale scores of 26.8, 54.1 and 72.6 assigned by the fibromyalgia patients to the moderate reference health state, their own health state and the mild reference health state respectively, may indicate a tendency to use the whole scale. In ankylosing spondylitis these baseline scores were 36.7, 80.0 and 76.5. Utilities are cardinal measures, reflecting not only the ranking of various health states relative to perfect health and death, but also the magnitude of the difference between these different health states.[23] If response spreading occurs, then the rating scale gives an indication of ordinal rankings and the intensity of the preferences, but it does not provide interval-scale utilities.

Risk Attitude

A second explanation for the significant difference between mean rating scale and mean standard gamble scores may be the patients' attitudes towards risk itself. Rating scale scores are measured under certainty, and do not capture the respondent's attitude towards risk. In contrast, the standard gamble approach incorporates the respondent's risk attitude, which may be risk averse, risk neutral or risk seeking. If subjects are not risk neutral, differences can be expected between rating scale and standard gamble values. If they are risk averse their indifference probability increases, thus increasing the utility of the health state being valued. Kahneman and Tversky have shown that people generally acted as if they were risk averse when choices were framed in terms of potential gains and as risk seeking when choices were framed in terms of potential losses.[31] According to their "prospect theory" in which an S-shaped value function is assumed, the displeasure of a loss is generally greater than the pleasure associated with an equivalent gain.[32] Although our gambles are mixed gambles with both a positive and negative potential outcome, the patients in our studies might have focussed more on the negative outcome of the gamble, i.e., the severe reference state. In that case the shape of the value function might have contributed to risk-averse behavior. As a result of this behavior, the standard gamble utilities would be biased upward compared to rating scale values.[17]

Standard gamble utilities may also be biased upward because people tend to over-weight sure outcomes relative to outcomes which are highly probable. This is called the certainty effect but it is also referred to as the Allais paradox.[17,31] Kahneman and Tversky's "prospect theory" assumes a decision weight function which over-weights small probabilities and under-weights moderate and high probabilities. If, in our studies, moderate and high probabilities were under-weighted, this might have contributed to the relative attractiveness of the sure outcome, even when the probability of gaining perfect health was rather high. Over-weighting a small chance of ending up in the severe reference state might have reinforced the attractiveness of the sure outcome. Moreover, the fact that patients knew from experience that they had been able to adapt to their illness before may have diminished both the severity of the health state being valued and the relative value of the therapeutic pay-off from treatment gambles.[33]

Cognitive and Emotional Factors

Other explanations for the difference between rating scale and standard gamble values are all related to the previous explanations. Such explanations might include differences in cognitive processes such as recalling and taking account of past events, life goals, family circumstances and the selection of reference points against which consequences are evaluated.[14]

Cognitive factors play an important role in Loomes and Sugden's alternative to expected utility

theory, called "regret theory".[34,35] According to that theory, the value a person assigns to a health state does not only depend on that health state but also on how that health state compares to the health state the person might have had if he or she had made a different choice. If what is obtained is better than what might have been, feelings of rejoice may increase the utility; if what comes is worse than what might have been, regret may reduce the utility. In standard gambles where patients are explicitly asked to make a choice, feelings of regret and rejoice may be anticipated, whereas such feelings are absent in the choiceless rating scale valuation process. Subjects may shy away from the gamble choice in the standard gamble because of regret aversion (regret may occur if they "lose" the gamble and end up with the worst outcome). By means of experiments, Loomes and colleagues have shown that regret theory is able to explain why observed preference reversals may have occurred.[36]
Finally, differences may arise in emotional reactions to past and future health states and events, such as more intense emotional reactions to bad outcomes when they are presented in gambles or more intense emotional reactions to probabilities of death if a family member or friend has died recently. The likelihood of such "recall reactions" is lower for the rating scale than for the more confrontational standard gamble method.

2. Reliability of the Rating Scale and the Standard Gamble

Test-retest reliability

The reproducibility of the rating scale and standard gamble methods has been scarcely reported in the literature. O'Connor reported a Pearson product moment correlation of 0.77 and 0.80 for the one week test-retest reliability of the rating scale and the standard gamble,[37] while Torrance reported product moment correlations of 0.49 and 0.53 for the one year test-retest reliability of these methods.[15] In our studies reproducibility was assessed by calculating 3 month intraclass correlation coefficients (ICC) for the values assigned to the reference health states. The results are given in table 3.

Table 3. Test-retest reliability: three months intraclass correlation coefficients

	Rating Scale		Standard Gamble	
Reference states	Fibromyalgia	Ank. Spondylitis	Fibromyalgia	Ank. Spondylitis
Mild	0.33	0.26	0.43	0.50
Moderate*	0.24	0.29	-	-
Severe**	-	-	0.70	0.65

* the moderate reference health state was not valued by means of the standard gamble
** the severe reference health state was not valued by means of the rating scale

Although the reproducibility of the standard gamble is somewhat higher than that of the rating scale, the ICCs of the scores assigned to the reference states are generally not very high. This may point at difficulties in valuing abstract, hypothetical health states that have never been

experienced. There are good reasons why it may be difficult to envision the well being associated with a hypothetical health state. One is the inevitable gap between imagination and the actual experience of a health state. Individuals may overestimate or underestimate their ability to accommodate or to cope with adversity.[13]
Although the reproducibility is not high, table 4 shows that - in spite of the occurrence of some slight but statistically significant changes in the preferences for the mild reference state in fibromyalgia and the severe reference state in ankylosing spondylitis - the mean scores are fairly stable over time. This stability may point at the usefulness of aggregated scores for group decision making.

Table 4. Mean (SE) utilities of the reference health states

	0 months	3 months	6/9 months*	p value** 0-3	p-value** 3-6/9
Fibromyalgia					
rating scale					
moderate	27.3 (1.55)	29.1 (1.68)	30.4 (1.54)	0.382	0.388
mild	72.6 (1.54)	79.7 (1.26)	76.8 (1.28)	0.000	0.035
standard gamble					
mild	0.83 (0.02)	0.85 (0.02)	0.85 (0.02)	0.366	0.851
severe	0.42 (0.04)	0.35 (0.04)	0.31 (0.04)	0.083	0.095
Ankylosing Spondylitis					
rating scale					
moderate	36.7 (1.14)	39.0 (1.28)	39.1 (1.37)	0.116	0.909
mild	76.5 (0.83)	77.3 (0.70)	77.5 (0.90)	0.346	0.888
standard gamble					
mild	0.84 (0.02)	0.85 (0.01)	0.86 (0.01)	0.777	0.453
severe	0.33 (0.03)	0.31 (0.03)	0.39 (0.03)	0.449	0.014

* The second followup measurement was scheduled at 6 months in fibromyalgia and at 9 months in ankylosing spondylitis.
** paired t-test

The observed slight changes in reported preferences may result from real changes in the patients' health states. However, in fibromyalgia no significant changes in patients' health states were found on a variety of clinical and quality of life outcome measures.[20] Moreover, there is some evidence to support the hypothesis that patients' valuations of states of health are not influenced by their actual health state.[38] For example, ankylosing spondylitis patients' health improved as measured on a number of outcomes, including a global assessment scale, but their valuations of the reference states remained relatively stable.
A patient's true preference may change over 3 months and the preference at one time may not be representative of the patient's long-term preference.[39] This hypothesis is supported by the higher reliabilities that were found when the first 15 fibromyalgia patients from the control (no intervention) group to report for the 6 months followup were asked to return for a 4 week test-

retest reliability assessment.[40] In this assessment, rating scale ICCs of 0.56 and 0.67 were found for the patient's own health state and the mild reference state respectively. The ICC of the standard gamble utilities was 0.66 for the patient's own health state, 0.74 for the mild and 0.94 for the severe reference state. The generally higher ICCs for the severe reference state are partly due to the fact that negative utilities for the severe reference state were recoded to zero.

Standard error of measurement

Another way of looking at the reliability of our utility measurements is by looking at the measurement error.[15,28,29] Each single patient's preference measurement contains some measurement error, which causes part of the variance among the scores. The standard error of measurement, which is the standard deviation of the measurement error, was calculated to be 0.13 for the standard gamble (scale from 0-1) in fibromyalgia and 0.11 for the standard gamble in ankylosing spondylitis. For the rating scale method (scale from 0-100), the standard deviation of the measurement error was 10 in fibromyalgia and 8 in ankylosing spondylitis. These figures suggest that both methods contain considerable measurement error, implying relative instability of an individual patient's preferences. This supports the notion that utilities may be less useful for individual decision-making than for group decision-making, as was indicated by the relative stability of preferences over time shown in table 4.

3. Internal consistency of the standard gamble

Basic reference gambles versus chained gambles

According to the axioms of expected utility theory, the outcomes of the gamble should not influence a patient's utility for a particular health state. The patients are supposed to adjust their indifference probability to allow for alterations in the gamble outcomes. At 9 months followup 80 consecutive ankylosing spondylitis patients were asked to value their own health state both in comparison to perfect health and death and in comparison to perfect health and the severe reference state. The latter was the first of a chained pair of gambles. According to expected utility theory, there should be no difference in utilities elicited by a basic reference gamble or a chained gamble. However, the mean utility value of 0.83 in case of the basic reference gamble was statistically significantly lower than the mean utility of 0.87 when the severe reference state was used as a gamble outcome in a chained pair of gambles (paired t-test; $p=0.018$). This is in accordance with earlier findings.[18,41] A majority of the patients (75%) assigned lower utilities in the basic reference gamble than in the chained gambles. Table 5 gives an example of this phenomenon for one patient.

Table 5. Difference between a chained and a basis reference gamble

SG question	Health state being valued	Gamble outcomes*	p**	U_{hi}***
2 (chained)	own	perfect health (1) severe ref. state (0.55)	0.75	0.89
5 (basic)	own	perfect health (1) death (0)	0.85	0.85

* Utilities for outcomes in parentheses ** Indifference probability for gamble
*** Utility for a patient's own health state

The fact that the chained gambles resulted in higher utilities than the basic reference gamble does not imply that patients took a smaller risk when death was replaced by the severe reference state. On the contrary, generally, patients took a greater risk in the gamble where the severe reference state was the worst outcome than in the gamble where death was the worst outcome. The eventually higher utilities in the chained gamble resulted from a combination of relatively small differences between the indifference points in the chained and basic reference gamble and the relatively high utilities assigned to the severe reference state (see example in table 5). This finding is in contrast to Llewellyn-Thomas et al. who found that raters were prepared to take a greater risk in gambles when death was the worst outcome.[18] This difference may be explained by the fact that Llewellyn-Thomas et al. elicited utilities from cancer patients, to whom death usually is a real risk, whereas death to fibromyalgia or ankylosing spondylitis patients is not or only a remote issue.

Since the standard gamble method seems to be susceptible to the characteristics of the worst outcome of the gamble, Hellinger and Llewellyn-Thomas conclude that this method is internally inconsistent.[17,18] However, when a change of focus or a change of reference point occurs as a result of a change in gamble outcomes, the preference shifts are not necessarily illogical.

Inconsistent Responses

Respondents are expected to provide preferences that are consistent with the natural underlying order of our health state descriptions. In other words, dominance violations should not occur. For the standard gamble as applied in our studies, dominance implies that when a patient's own health state is compared to the mild reference state and all 6 dimensions indicate a better (worse) or equal functional level, the utility the patient assigns to his own health state should be higher (lower) or equal to the utility of the mild reference state. However, 7 fibromyalgia patients - each only once - did not provide preferences in accordance with this expectation. In ankylosing spondylitis dominance was violated by 17 patients on the standard gamble. On the total number of questions this number of dominance violations is rather low. Moreover, in 17 of the 24 (71%) inconsistent answers, the difference between the utilities of the two compared health states was smaller than the standard deviation of the measurement error reported in the previous section. Thus, most of the inconsistent responses fell within the bounds of the measurement error of 0.13 in fibromyalgia and 0.11 in ankylosing spondylitis.

In our studies every patient preferred his or her baseline health state to the severe reference state. Therefore, the valuation of a patient's followup health state in a gamble with perfect health and the severe reference state should result in a higher or equal indifference probability than the valuation of the same health state in a gamble with perfect health and the patient's

baseline health state. When every followup measurement was checked, we found that 6 (9%) fibromyalgia patients and 6 (4%) ankylosing spondylitis patients violated this rule once. It would also be expected that, if death is regarded as worse (better) than or equal to the severe reference state, the indifference probability of the patient's own health state when compared to perfect health and death should be higher (lower) than or equal to its indifference probability when compared to perfect health and the severe reference state. This was checked for the 80 ankylosing spondylitis patients who were asked the additional fifth standard gamble question. Twelve of them (15%) gave an inconsistent response. Again the majority (18) of these 24 inconsistent responses fell within the bounds of the measurement error.
Overall, about 21% of all patients gave an inconsistent response. By means of a logistic regression, we enquired whether the patients who gave inconsistent responses were somehow different from the patients who did not. In ankylosing spondylitis we found that, when controlling for the influence of all other variables, males were more likely to give an inconsistent response than females. Otherwise, no differences with respect to age, duration of illness, education and marital status was found. We have no explanation for this gender related difference. Perhaps it is due to chance.
The inconsistencies described above are perhaps due to the fact that the health state descriptions cover too many dimensions for some people to include every dimension in their overall valuation. Concentration on just one or two dimensions that are considered important can lead to inconsistent responses. Some patients assigned higher preferences to their own followup health state when their baseline health state was used as the worst outcome of the gamble than when the severe reference state was used. This may be more an indication of appreciation for even a small improvement than of a change in preference.

Extreme Risk Averse Behavior: assigning a utility of 0.95 to each health state
Eight out of 85 (9%) fibromyalgia patients reached the same indifference probability of 0.95 for all health states being valued in the first 3 standard gamble questions on at least one of the measurement times. Eleven out of 144 (8%) ankylosing spondylitis patients assigned a value of 0.95 to all 3 health states at least once. One fibromyalgia patient and one ankylosing spondylitis patient did this consistently at baseline and each followup measurement. When controlling for other patient characteristics, fibromyalgia patients in which this phenomenon was found tended to be somewhat older than the patients in which this was not found. However this difference was not statistically significant (logistic regression; Wald statistic=2.975; p=0.08).
Apparently these patients were never willing to take a larger than 10% risk of getting the worst outcome, irrespective of the severity of illness in the health state being valued. This behavior can be explained by a general aversion to gambling.[13] Such a reluctance to comply with the standard gamble questions reflects either reluctance to face the reality of the decision problem, reluctance to bear decision-making responsibility or inability to grasp hypothetical, unrepresentative experiments presented in a necessarily simple and abstract way. It may have been too difficult for these patients to imagine well being associated with a hypothetical health state or they may have underestimated their ability to cope with the severe reference state.
The difference between the individual and the group perspective also sheds some light on this phenomenon.[42] Risk might seem higher from an individual perspective than from an aggregate, group perspective. A single patient will either become perfectly healthy or severely ill (or die in case of standard gamble question 3), and will never receive the average burden or average benefit. Therefore the expected utility may seem very abstract to a single patient in the 1-game setting, but not in the 100 game setting.[43]

Finally, the probability steps of 10%, from 100/0 to 90/10 etc., may have been too large. Changes of 5% (100/0, 95/5, 90/10 etc.) or even 1% (100/0, 99/1, 98/2 etc.) could have produced a difference in utilities between the different health states. However, findings by Kahneman and Tversky suggest that probabilities of less than 0.1 and greater than 0.9 are difficult for people to handle.[31]

4. A Health State worse than Death

In standard gamble question 3 the severe reference health state is valued against perfect health and death. Some 41 of the 85 fibromyalgia patients (48%) and 78 of 144 (54%) ankylosing spondylitis patients indicated at least once that the severe reference state was worse than death. This means that they were prepared to accept a 100% risk of dying to avoid this state. Eighteen fibromyalgia patients and 20 ankylosing spondylitis patients preferred death to the severe reference state at all 3 measurements. When asked explicitly, all these patients confirmed that they would rather die than live in the severe reference state.
As mentioned before, when patients indicated that the severe reference state was worse than death, a utility of zero was assigned to this state. To actually measure the magnitude of a negative utility, Torrance has suggested a slight modification of the usual standard gamble question.[10] In this modified question, patients are offered a choice between a sure death or a gamble with chance p of perfect health and chance 1-p of living irreversibly in the severe reference state. This question is presented by asking patients to imagine that they suffer a rapidly progressing terminal disease which - if left unattended - will lead to death. However, if treated there is a chance of gaining perfect health or of becoming like the severe reference health state for the remainder of their life. The utility of the severe reference state is calculated, from the indifference probability, as -p/(1-p).
At the 3 and 9 months followups all the ankylosing spondylitis patients who indicated that the severe reference state seemed worse than death (51 at 3 months and 36 at 9 months) were asked this standard gamble question. At 3 months followup, the 5% trimmed mean utility of the severe reference state was -0.16 with a 5% trimmed standard deviation of 0.19; at 9 months followup the 5% trimmed mean utility of the severe reference state was -0.18 with a 5% trimmed standard deviation of 0.34. The mean and standard deviation were trimmed 5% because the distribution of the answers to standard gamble question 3 is heavily skewed to the left. The standard gamble to assess the magnitude of a negative utility can result in utilities smaller than -1, thus causing the upper end of the utility scale (from 0 to 1) to be shorter than the lower end of the scale (from 0 to -19 or less, depending on the size of the probability increments in the measurement instrument). At 3 months followup 1 patient assigned a utility of -3.00 and at 9 months followup 3 patients assigned a utility lower than -1 (-1.86, -3.00 and -19.00) to the severe reference state.
The finding that for so many patients death was not the worst imaginable outcome has led some authors to conclude that death is not the logical zero point for a utility scale.[44] Although we certainly recognise the existence of health states valued worse than death, the allowance of negative utilities is problematic. Furthermore, the descriptive validity of negative utilities may be questioned because probably few people would act in accordance with their statement that they would be better off dead. The negative utility for a severe state of illness can change into a positive one when actually experiencing that state. Moreover, the occurrence of utilities below -1 is problematic too. Based on the large number of negative utilities for the severe reference

state used as the anchor point in our gambles and the observed magnitude of these utilities (some of which were much lower than -1), we conclude that, for the present, it is more convenient to use death as the anchor point in the gambles. Moreover, this increases the comparability of utilities measured in different patient groups and in different studies.

CONCLUSIONS AND DISCUSSION

Table 6 provides an overview of the most important findings, possible explanations and preliminary implications reported in the previous sections.

In the introduction of this article 2 different approaches to utility measurement in clinical trial settings were distinguished. One can either use prepackaged systems (e.g. the Quality of Wellbeing scale, Rosser's Disability Distress scale or the recently developed EuroQol[45]) or directly elicit preferences from patients. When using prepackaged systems one needs to acquaint oneself with the method on which the underlying utilities were based. As has been reported before and reconfirmed in our studies, standard gamble preference scores for a particular health state are significantly higher than rating scale preference scores for that same health state. Hence, since the Quality of Wellbeing scale, the EuroQol and Rosser's Disability Distress scale are based on rating scales, these prepackaged instruments would be expected to produce lower scores than the McMaster Health Utility Index which is based on the standard gamble. When measuring preferences directly one again has the choice between preference rankings or choice-based methods. Moreover, if one decides to use the standard gamble one can also choose between various ways of taking the measurements: between basic reference gambles or chained gambles and between absolute or relative change questions. It was found in our studies that these different versions of the same method resulted in statistically significant utility differences. Chained gambles resulted in higher utilities than the basic reference gamble and absolute changes in utilities were smaller than relative changes measured directly in comparison to the baseline health state.

Table 6. Summary table.

Most important findings	Possible explanations	Implications
1. Comparing RS and SG utilities*		
- SG utilities significantly higher than RS utilities.	- response spreading on the RS - risk attitude incorporated in the SG - cognitive and emotional factors (eg. regret theory)	- prepackaged utility measurement instruments based on the SG are bound to produce higher utilities - costs per QALY will be sensitive to the method of utility measurement used
2. Reliability of RS and SG*		
- The ICC's of the reference health states were higher for the SG than for the RS, but low for both methods. - Mean scores of both methods were rather stable. - The standard error of measurement was about 0.12 for the SG (0-1 scale) and about 9 for the RS (0-100 scale).	- instability of intra-individual valuations due to difficulties in valuing hypothetical states, but stability of mean values at the group level - substantial measurement error inherent to utility measurement - precision decreases as choice is introduced	- limited use of utilities for individual decision making in clinical decision analysis - more confidence in the use of utilities for program evaluation - need to undertake repeated measurements
3. Internal consistency of the standard gamble		
- Chained gambles resulted in higher utilities than basic reference gambles. - About 21% of all patients gave at least 1 inconsistent response. - About 70% of the inconsistent responses fell within the bounds of the standard error of measurement. - Almost 10% of all patients assigned a utility of 0.95 to three very different health states; they were not willing to take any risk.	- outcomes of the gamble influence an individual patient's utility for a particular health state - inconsistent responses are to a large extend due to measurement error - general aversion to gambling - inability to understand hypothetical, abstract questions - risk may seem higher from an individual than from a group perspective - probability increments of 10% are too large - difficulty of SG method - health state descriptions cover too many dimensions	- costs per QALY depend on the outcomes used in the SG - repeated measurements may reduce inconsistency - SG not suited for everyone
4. A health state worse than death		
- About 50% of the patients valued the severe reference state used in the chained gamble as worse than death on at least 1 occasion. - Negative utilities smaller than -1 occurred.	- recognition of the existence of health states worse than death - low validity of the response	- an appropriate way to incorporate negative utilities is needed - searching for alternative ways to calculate negative utilities - for the time being, negative utilities may be avoided for practical reasons

* RS = Rating Scale; SG = Standard Gamble ;**ICC = Intraclass Correlation Coefficient

Utilities are proposed as a decision aid in 2 different contexts: 1) where choices have to be made between alternative therapies for the same individual,[26] and 2) where choices have to be made between alternative ways of allocating limited resources among different health care activities serving the same or different patient groups.[46] There is cause for misgivings regarding the use of utilities for clinical decision analysis in the first context. For example, a number of inconsistencies were found in the responses of single patients to different standard gamble questions. It is likely that these reflect the underlying measurement error in taking a single preference measure. Our study indicates that the standard deviation of the measurement error ranged from 0.11 to 0.13 for the standard gamble and from 8 to 10 for the rating scale. Overall, more than 70% of the inconsistent responses fell within the bounds of the measurement error. The relatively poor stability of measurements from an individual patient limits the use of utilities based on a single measurement only for individual decision-making. This increases the need to undertake the measurements repeatedly to average out the measurement noise within the individual.

Even though the 3 months test-retest reliability was not very high, the relative stability of the mean utilities over time on the group level gives some confidence in the use of utilities for program evaluation in cost-utility analysis. However, since generally only a single estimate of utility is used in cost-utility analysis, this analysis may be very sensitive to the method used to elicit utilities. Recently Hornberger has shown that the effect of different methods on the final cost-utility may be considerable.[47] It is as yet premature to suggest a preferred method of utility measurement. Neither the rating scale nor the standard gamble method seems superior. As to the rating scale, there remain fundamental doubts about the interval properties of the scale.[48] Moreover, the repeatability of this method was found to be lower than that of the standard gamble. As to the standard gamble, it would be interesting to determine to what extent the observed difference between basic reference gambles and chained gambles influences the final results of cost-utility analysis.

Many of the inconsistencies and responses that seem to violate expected utility theory which we found may be associated with the description of the severe reference state used as the worst outcome in the first and as the certain outcome in the second of a chained pair of standard gambles. The health state of a severely ill person may be so hard to imagine that patients put all their effort into understanding it, instead of paying attention to what is really being asked. This possibility is supported by the fact that a number of patients changed their opinions as to whether the severe state was better or worse than death. Overall about 50% of the patients indicated at least once that the severe reference state was worse than death. When attempting to measure how much worse than death, a small number of highly negative utilities occurred that greatly influenced the mean utility. These findings and the difficulties of handling extreme negative utilities may argue in favor of using death as the worst outcome of the gamble. Death may be a more imaginable zero point to anchor the scale.

As preference scores elicited by the standard gamble seem to be susceptible to the way questions are presented and the endpoints used, doubts may be raised about the validity of expected utility theory. However, since there is no evidence that less formal procedures to guide individual therapy decisions and resource allocation decisions are any less susceptible to the effects of different methods of presentation and various biases, it is not appropriate to reject utilities as useful outcome measures. Moreover, as Torrance and Feeny point out, expected utility theory may be regarded as normative as opposed to behavioral.[25] This theory describes how an individual should behave if he or she wished to act rationally in order to maximize the expected utility. It does not describe how an individual actually makes a decision under

uncertainty. Perhaps patients would have made more rational responses if they had been better informed about the meaning of their responses and the consequences of their choices. Furthermore, many of the observed inconsistencies or preferences that do not seem to fit expected utility theory are not necessarily illogical. Some of the alternatives to expected utility theory, such as prospect theory and regret theory, may help to explain several of the seemingly inconsistent answers from a more behavioral perspective. The challenge is to explore the potential contribution of such theories to utility measurements in health care decision settings.

Acknowledgements
We would like to thank the anonymous reviewer, George Torrance, Han Bleichrodt, Grant Rhodes, Silvia Evers, Margreet Janssen, Linda Heurman and Mariëlle Goossens for their valuable comments on an earlier draft of this paper.

REFERENCES

1 Feeny D, Labelle R, Torrance GW: Integrating economic evaluations and quality of life assessments. In: Spilker B ed. Quality of Life Assessment in Clinical Trials. New York: Raven Press, 1990.

2 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.

3 Gerard K: Cost-utility in practice: A policy maker's guide to the state of the art. Health Policy 1992;21:249.

4 Rosser R, Kind P: A scale of valuations of states of illness: is there a social consensus? Int J Epidemiology 1978;7:347-58.

5 Kaplan RM, Bush JW: Health status: types of validity and the index of well-being. Health Serv Res 1976;478-507.

6 Torrance GW, Boyle MH, Horwood SP: Application of multi-attribute utility theory to measure social preferences for health states. Operations Res 1982;30:1043-69.

7 Bombardier C, Ware J, Russell IJ, Larson M, Chalmers A, Read JL: Auranofin therapy and quality of life in patients with rheumatoid arthritis. Results of a multicenter trial. Am J Med 1986;81:565-78.

8 Thompson MS, Read JL, Hutchings HC, Paterson M, Harris ED: The cost effectiveness of auranofin: results of a randomized clinical trial. J Rheumatol 1988;15:35-42.

9 Mehrez A, Gafni A: Quality-adjusted life years, utility theory, and health-years equivalents. Med Decis Making 1989;9:142-9.

10 Torrance GW: Measurement of health state utilities for economic appraisal. A Review. J Health Econ 1986;5:1-30.

11 Neumann Von J, Morgenstern O: Theory of games and economic behavior, Princeton: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).

12a Froberg DG, Kane RL: Methodology for measuring health-state preferences-I: Measurement strategies. J Clin Epidemiol 1989;42:345-54.

12b Froberg DG, Kane RL: Methodology for measuring health-state preferences-II: Scaling methods. J Clin Epidemiol 1989;42:459-71.

12c Froberg DG, Kane RL: Methodology for measuring health-state preferences-III: Population and context effects. J Clin Epidemiol 1989;42:585-92.

12d Froberg DG, Kane RL: Methodology for measuring health-state preferences-IV: Progress and a research agenda. J Clin Epidemiol 1989;42:675-85.

13 Mulley AG: Assessing Patients' Utilities. Can the ends justify the means? Med Care 1989;27:S269-S81.

14 Read JL, Quinn RJ, Berwick DM, et al: Preferences for health outcomes: comparisons of assessment methods. Med Decis Making 1984;4:315-29.

15 Torrance GW: Social preferences for health states: an empirical evaluation of three measurement techniques. Socioecon Planning Sci 1976;10:129-36.

16 Wolfson AD, Sinclair AJ, Bombardier C, McGreer: Preference measurement for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL and Kane RA (eds.). Values and Long-Term Care. Lexington Books, D.C. Heath and Company, Lexington, 1982.

17 Hellinger FJ: Expected utility theory and risky choices with health outcomes. Med Care 1989; 27:273-9.

18 Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF: The measurement of patients' values in medicine. Med Decis Making 1982;2:449-62.

19 Hidding A, van der Linden S, Boers M, et al: Is group physical therapy superior to individualized therapy in ankylosing spondylitis. Arthritis Care Res 1993;6:117-25.
20 Santen van-Hoeufft M, Bolwijn P, Kleijnen J, et al: Is low or high impact fitness beneficial in fibromyalgia patients. Results of 2 randomized controlled trials. (in preparation).
21 Wolfe F, Smythe HA, Yunus MB, et al: The American College of Rheumatology 1990. Criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. Arthritis Rheum 1990;33:160-72.
22 Linden van der S, Valkenburg HA, Cats A: Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.
23 Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991;(suppl)12:118-28.
24 Bakker CH, Rutten-van Mölken M, Van Doorslaer E, Bennett K, Van der Linden S: Health related utility measurement in rheumatology: an introduction. Patient Education and Counseling 1993;20:145-52.
25 Torrance GW, Feeny D: Utilities and quality-adjusted life years. Int J Technol Assess Health Care 1989;5:559-75.
26 Weinstein MC, Fineberg HV: Clinical Decision Analysis. Philadelphia: W.B. Saunders Company, 1980.
27 Norusis/SPSS INC. Advanced Statistics SPSS/PC+. Chicago, 1988.
28 Deyo RA, Diehr P, Patrick DL: Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clin Trials 1991;12:142S-58S, 1991.
29 Streiner DL, Norman GR: Health Measurement Scales. A practical guide to their development and use. Oxford University Press, 1989;83-9.
30 Kaplan RM, Bush JW, Berry CC: Health status index: category rating versus magnitude estimation for measuring levels of well being. Med Care 1979;17:501-25.
31 Kahneman D, Tversky A: Prospect theory: An analysis of decision under risk. Econometrica 1979;47:263-91.
32 Tversky A, Kahneman D: The framing of decisions and the psychology of choice. Science 1981;211:453-8.
33 O'Brien BJ, Elswood J, Calin A: Willingness to accept risk in the treatment of rheumatic disease. J Epidemiol Community Health 1990;44:249-52.
34 Loomes G, Sugden R: Regret theory: an alternative theory of rational choice under uncertainty. Econ J 1982;92:805.
35 Loomes G, Sugden R: Some implications of the more general form of regret theory. J Economic Theory 1987;41:270.
36 Loomes G, Starmer C, Sugden: Observing violations of transitivity by experimental methods. Econometrica 1991;59:425-439
37 O'Connor AM, Boyd NF, Till JE: Methodological problems in assessing preferences for alternative therapies in oncology: The influence of preference elicitation technique, position order and test-retest error on the preferences for alternative cancer drug therapies. Nursing research: science for quality care; Proc 10th National Nursing Research Conference. Toronto: University of Toronto, 1985:49-58.
38 Llewellyn-Thomas HA, Sutherland HJ, Tritchler DL, et al: Benign and malignant breast disease: The relationship between women's health status and health values. Med Decis Making 1991;11:180-88.
39 Christensen-Szalanski JJJ: Discount functions and the measurement of patients' values. Women's decisions during childbirth. Med Decis Making 1984;4:46.
40 Bakker C, Rutten-van Mölken M, Van Doorslaer E, Bennett K, Van der Linden S: Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. J Rheumatol 1994;21:269-74.

41 Bleichrodt H: Testing the validity of expected utility theory in health state valuation: Some experimental results. Institute for Medical Technology Assessment paper no 93.23, Rotterdam, Erasmus University of Rotterdam, 1993.
42 Redelmeier DA, Tversky A: Discrepancy between medical decisions for individual patients and for groups. N Engl J Med 1990;322:1162-4.
43 Asch DA: Choices for individual patients vs groups. Letter to the editor. N Engl J Med 1990; 323:922.
44 Haig THB, Scott DA, Wickett LI: The rational zero point for an illness index with ratio properties. Med Care 1986;24:113-24.
45 The EuroQol Group. A new facility for the measurement of health related quality of life. Health Policy 1990;16:199-208.
46 Laupacis A, Feeny D, Detsky AS, Tugwell PX: How attractive does a new technology have to be to warrant adoption and utilization? Can Med Assoc J 1992;146:473-81.
47 Hornberger JC, Redelmeier DA, Petersen J: Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. J Clin Epidemiol 1992;45:505-12.
48 Nord E: The validity of a visual analogue scale in determining social utility weights for health states. Int J Health Planning and Management 1991;6:234.

8

PROBLEM ELICITATION TO ASSESS PATIENT PRIORITIES IN ANKYLOSING SPONDYLITIS AND FIBROMYALGIA.

Carla Bakker, Sjef van der Linden, Marijke van Santen-Hoeufft, Paulien Bolwijn, Alita Hidding

Department of Internal Medicine, Division of Rheumatology, University of Limburg, Maastricht, The Netherlands

Accepted for publication in the Journal of Rheumatology

PROBLEM ELICITATION TO ASSESS PATIENT PRIORITIES IN ANKYLOSING SPONDYLITIS AND FIBROMYALGIA.

ABSTRACT

Objective. To elicit patient priorities as outcome measures in ankylosing spondylitis (AS) and fibromyalgia (FMS); to relate these measures to other outcomes; to assess construct validity and sensitivity to change of the problem elicitation technique (PET) questionnaire.

Methods. 134 Patients with AS were randomly allocated to weekly sessions of group physical therapy or daily exercises at home, whereas 73 patients with FMS were randomized into one of three groups (low-impact fitness, biofeedback, controls). The PET questionnaire was applied by trained interviewers at baseline and at each 6 (FMS) and 9 (AS) months' followup. A PET score was calculated at each assessment. Construct validity of the PET was assessed by correlation and multiple regression of baseline values with other disease outcomes (pain, stiffness, patient's global assessment, Sickness Impact Profile (SIP), Health Assessment Questionnaire (HAQ), Arthritis Impact Measurement Scale (AIMS), patient-utilities). Sensitivity to change of PET was assessed against changes in these outcomes and by comparing the efficiency of the PET with other outcomes.

Results. Patients with FMS identified more problems (mean 6.8) than patients with AS (mean 4.4). Moreover, more often AS than FMS patients could not identify any problem at baseline (10% compared to 1%). The PET score improved from 14.9 to 11.3 ($p=0.0001$) in AS patients but did not change from 21.8 to 21.1 ($p=0.24$) in FMS patients. Construct validity testing of the PET score showed statistically significant ($p<0.05$) correlations with AIMS, utilities, SIP, HAQ, pain, stiffness, and patient's global health in both AS and FMS patients (r's varying from 0.22 to 0.66). By multiple regression pain explained 29% of the variance in PET scores among patients with AS. In FMS patient's global assessment accounted for 39% of total variance of PET scores, whereas pain explained another 15%. Changes in PET scores correlated significantly ($p<0.05$) with changes in AIMS, utilities, pain, stiffness, and patient's global health in both AS and FMS (r's varying from 0.22 to 0.51). Some 6% of the variance in changes in PET scores was explained by changes in pain in AS patients and for 35% by changes in pain and subjective health in FMS patients. Assessment of sensitivity to change revealed that efficiency of the PET score was 0.6 in AS patients and 0.09 in FMS patients. Compared to other outcomes this was reasonable in AS patients but low in FMS patients.

Conclusion. Obtaining patient priorities was generally feasible. In both AS and FMS patients construct validity of the PET questionnaire was satisfactory. The PET was much more sensitive to change in AS patients than in FMS patients.

INTRODUCTION

In rheumatology interest is growing in measuring patient's preferences[1], because these measures may be more responsive to relevant improvement than traditional questionnaires.[2] The Problem Elicitation Technique (PET), a preference questionnaire dealing with disabilities, evolved from the McMaster Toronto Arthritis Rheumatism (MACTAR) patient preference disability questionnaire.[1,2] PET and MACTAR were developed as outcome measures to evaluate treatment effects in patients with rheumatoid arthritis. Small but clinically important changes in function may be detected by the PET or MACTAR because these instruments focus on activities that are limited by the disease and are considered important by the patient.[2] The use of patient focussed responsive instruments may reduce the sample size needed in clinical trials.[2]
In patients with ankylosing spondylitis (AS) or fibromyalgia (FMS) we were able to elicit patient preferences by the PET questionnaire.[3,4] In this chapter we concentrate on construct validity and discriminant validity (or sensitivity to change) of the PET relative to improvements in other outcomes.

PATIENTS AND METHODS

Study groups
The PET questionnaire was applied in 2 clinical trials.[3,4] In the **first** study 144 patients with AS (modified New York criteria)[5] were randomized to daily exercises at home (n=75) or weekly sessions of physical therapy in groups in addition to daily home exercises (n=69) (3 hours physical training, sports and hydrotherapy at each weekly session). Ten patients dropped out during the 9 month followup [9 (12%) in the "at home" group and 1 (1%) from the group physical therapy (exact 2 tailed p value of 0.018)]. Reasons for dropping out were: moved (1), pregnant (1), spinal surgery (1), cardiac or lung disease (2), inability to exercise (4), too busy (1). The baseline characteristics of those who dropped out did not differ significantly ($p>0.05$) from the other patients, whereas patients who received group physical therapy did not differ at baseline from patients who exercised at home.[3]
In the **second** study, 85 patients with FMS (female sex, age 18-60 years, ACR criteria[6]) were randomized to either low-impact fitness training (n=35) (supervised aerobic and stretching exercises for 60 minutes, twice weekly, during 6 months), biofeedback (n=31) (20 minutes supervised relaxation training twice a week, during 2 months), or controls (n=19).[4,7] The PET questionnaire could not be applied at baseline to 3 patients with FMS. Eleven patients dropped out during the 6 month followup: 5 in the fitness group; 5 in the biofeedback group; 1 from the control group. Their baseline characteristics did not differ significantly ($p>0.05$) from the other patients (data not shown), whereas the baseline characteristics among the 3 groups did not differ except for age. Patients in the fitness group were on average older than the controls (44.9 years compared to 40.1 years) ($p=0.05$).[7]
In patients with AS or FMS the mean (SD) age was 43 (10) and 44 (8) years with a mean (SD) duration of disease of 7 (7) and 12 (10) years respectively. In the AS study 21% of the patients

were females. In the FMS study 100% of the patients were female.

The PET questionnaire
The PET questionnaire[8] was translated into Dutch by a qualified native translator and pre-tested in a small sample of volunteers. However, we did not apply the full set of guidelines to preserve equivalence in cross-cultural adaptation of health related quality of life measures as recommended in a recent paper.[9] The PET questionnaire was administered in a standardized way by trained interviewers at baseline and at each 6 (FMS) and 9 (AS) months followup. The PET questionnaire will be explained in 4 steps: identifying problems, impact of problems, summarizing problems into a PET score, and assessing patient's health as his (her) subjective health score.

- Identifying problems.

At baseline patients are asked to consider daily routine problems they have been experiencing during the last week as a result of their disease. Then they are asked to identify those problems that are most important to them and that they would like to see improved. Once the patient has finished identifying problems spontaneously, the interviewer read a series of 'probes' to assist the patient. These probes are open-ended questions covering 9 broad areas of function: self-care, mobility, role activities, leisure activities, communication, social interaction, sleep and rest, emotional health, and appearance. Patients are allowed to identify up to 15 problems.

- Impact from patient's point of view.

For each problem the patient was asked to fill out a form that shows 4 Likert scales (Figure 1). For problems related to the areas of self-care, mobility, role activities, leisure activities, communication and social interaction the **level of difficulty** on the first scale is marked, with '0' equal to 'without any difficulty' and '7' equal to 'unable to do'. For problems related to sleep and rest and emotional health the **degree of severity** is marked on the second scale, with '0' equal to 'none' and '7' equal to 'severe'. For problems related to appearance the **frequency** is marked on the third scale with '0' equal to 'never' and '7' equal to 'always'. For each problem the **importance** is also scored on the fourth scale with '0' equal to 'least important' and '7' equal to 'most important' (Figure 1). Once the patient has completed this form for each problem he or she was asked to reconsider the order of importance of all problems with respect to each other. Patients are allowed to change scores if so wished.

- Summarizing as a PET score.

A PET score was calculated by multiplying the difficulty (or severity or frequency) score by the importance score. Next these results were summed up for all problems and divided by the number of problems for each patient. Therefore, the maximum score is 49 and the minimal score is zero. A higher PET score indicates a higher degree of perceived disability. In our study the number of problems was kept constant at followup visits.

• Subjective health score.
Patients were asked to consider their overall health over the last week and indicate this on a numerical rating scale (1-10) with 'worst possible health' (1) at the left end and 'perfect health' (10) at the right end of the scale.

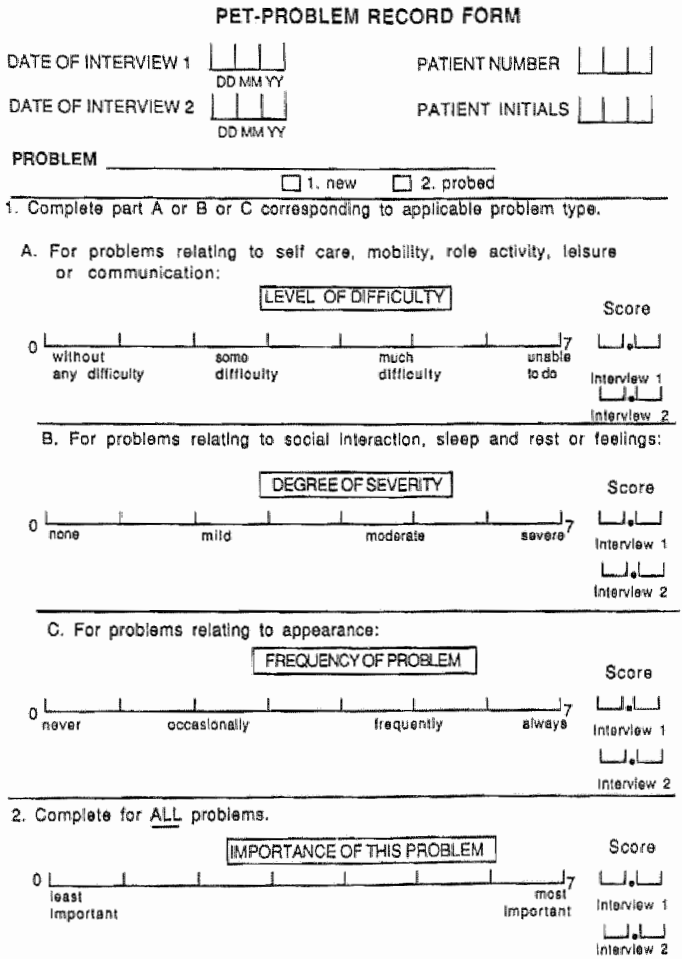


Figure 1. PET-problem record form

At *followup* for each problem patients were asked to complete the same forms as they had completed at baseline. Therefore, patients noticed their previous scores for each problem regarding difficulty, severity, frequency, importance and also their previous subjective health score.

Other outcome measures

In *both* studies the following measures were applied: the Sickness Impact Profile (SIP)[10]; the Dutch Arthritis Impact Measurement Scale (Dutch-AIMS)[11]; the Health Assessment Questionnaire (HAQ-S in AS and mHAQ in FMS)[12,13]; pain (on a 100 mm visual analogue scale (VAS) with 0 equal to 'no pain' and 100 equal to 'most severe pain imaginable'); stiffness (indicated by patients with AS on a 100 mm horizontal VAS with 0 equal to 'no stiffness' and 100 equal to 'worst stiffness I can imagine'; patients with FMS reported duration of stiffness in minutes). In addition, the Maastricht Utility Measurement Questionnaire (MUMQ) to assess patient's utilities was applied to all patients with FMS.[7] In the AS study the results of the utility measurement were confined to the 59 patients with AS, to whom the MUMQ was applied by the same interviewer at baseline and after 9 months' followup.[14]

In *ankylosing spondylitis* we assessed **also** the functional and the articular index[15]; enthesis index[16]; chest expansion, cervical rotation, thoracolumbar flexion and extension[17]; and physical fitness.[3] At baseline physicians assessed the activity of the disease on a 5 point scale with 1 equal to 'low disease activity' and 5 equal to 'high disease activity'. Assessment of improvement (or deterioration) was assessed by asking the patient to describe his or her perceived **change** in general functioning on a 10 cm horizontal VAS
(-5 = maximum worsening, 0 = no change, +5 = maximum improvement).

In addition, in *fibromyalgia* patients we elicited global health during the last week on a 0-10 numerical rating scale (NRS) with 0 equal to 'very bad' and 10 equal to 'very good health'.

Statistics and analysis

The findings of the PET questionnaire were summarized by mean and standard deviation (SD). For both the AS and FMS study baseline differences between intervention groups regarding number of problems, PET scores and subjective health scores were analyzed by unpaired student's t-test. At followup groups were compared for median improvement (or change) scores by Kruskal Wallis test. Within each intervention group changes were assessed by Wilcoxon signed rank test.

Construct validity[18] of PET was tested by the Spearman rank correlation coefficients between *baseline* PET values, and *baseline* scores for SIP, AIMS, HAQ, pain, stiffness, MUMQ, subjective health score, and the other measures for AS or FMS. Improvements in PET score were supposed to be associated with improvements in these variables. Therefore, p values were tested unilaterally (1-sidedly) at an α level of 0.05. Multiple regression analysis (forward selection) was performed with these measures and subjective health score, age and disease duration as independent variables and PET score as dependent variable. Independent variables with skewed distributions were analyzed as ln(var+1), the natural logarithm of one plus the variable.[19,20] The statistical analyses were performed with the SAS computer program [21] on an Olivetti M290S personal computer.

Discriminant validity (sensitivity to change)[18] was tested unilaterally by Spearman rank correlation coefficients between the *changes* (followup minus baseline) in PET scores and the *changes* (followup minus baseline) in other outcome measures and also by multiple regression analysis (forward selection) with *changes* in PET score as dependent variable and the trial intervention, *changes* in subjective health score and other health status outcome measures as independent variables. Both dependent and independent variables in the multiple regression analysis were transformed by ln(var+constant), the natural logarithm of a constant just below the minimum of (change in) the variable plus (change in) the variable.

Sensitivity to change was also assessed by calculation of the efficiency (E)(the mean change of

the measure divided by the standard deviation of change) as suggested elsewhere.[22]

RESULTS

For both patients with AS or FMS the duration of the interview was about 15 to 20 minutes at baseline, whereas at followup it was about 5 to 10 minutes. No interviews were broken off.

Number of problems

The number of problems was a mean of 4.4 (SD, 2.5) for patients with AS and 6.8 (SD, 2.5) for patients with FMS. Interestingly, 15 (10%) patients with AS and 1 (1%) patient with FMS could not identify any problem. Of the 129 patients with AS who indicated problems, 34 (26%) mentioned more than 5 problems compared with 56 (68%) of the 81 patients with FMS. Table 1 shows the categories of these problems.

At baseline the mean (SD) PET score and mean (SD) subjective health score were 14.9 (8.3) and 7.1 (1.6) respectively for all patients with AS, and 21.8 (7.6) and 6.1 (1.5) for all patients with FMS.

Table 1. PET questionnaire: 9 categories of problems

	AS	FMS
Number of patients	129	81
Self-care	6%	4%
Mobility	47%	28%
Role activities	3%	16%
Leisure activities	6%	8%
Communication	2%	3%
Social interaction	1%	1%
Sleep and rest	11%	16%
Emotional health	20%	22%
Appearance	4%	2%

Changes in PET scores at followup

For each disease the number of problems mentioned at baseline, as well as mean PET scores and mean subjective health scores did not differ between the intervention groups (data not shown).

- *Ankylosing Spondylitis patients*

The PET score improved significantly from a mean (SD) of 14.9 (8.3) at baseline to 11.3 (8.4) at 9 months followup [improvement:-3.6 (6.0); Wilcoxon signed rank test: $p=0.0001$]. The subjective health score improved slightly but statistically significant from a mean (SD) of 7.1 (1.6) to 7.3 (1.5) [change: 0.2 (1.2); Wilcoxon signed rank test: $p=0.04$]. The mean (SD) improvement in PET score was -3.9 (5.2) for the group physical therapy and -3.1 (6.7) for the "at home" group. The mean (SD) improvement in subjective health score was 0.4 (1.1) for the group physical therapy and 0.05 (1.3) for the "at home" group. These improvements did not

differ significantly between both AS intervention groups (Mann-Whitney test: p=0.16 and p=0.11 respectively).

- *Fibromyalgia patients*

The PET score was similar between 21.8 (7.6) at baseline to 21.1 (9.8) at 6 months followup [change: -0.7 (7.4); Wilcoxon signed rank test: p=0.24]. The subjective health score deteriorated slightly but statistically significant from a mean (SD) of 6.0 (1.6) to 5.7 (7.4) [change -0.4 (1.4); Wilcoxon signed rank test: p=0.03]. The PET score improved with a mean (SD) of -2.1 (6.8) in the fitness group, deteriorated in the biofeedback group and in the controls with a mean (SD) of 0.2 (8.9) and 0.5 (5.7) respectively. The mean (SD) subjective health score did not change in the fitness group [0.01 (1.2)], deteriorated in the biofeedback group and in the controls [-0.7 (1.4) and -0.5 (1.7) respectively]. The changes in PET score and subjective health score did not differ significantly across the 3 treatment groups (Kruskal Wallis test: p=0.32 and p=0.24 respectively).

Construct validity of the PET questionnaire

- *Ankylosing spondylitis patients*

PET scores correlated significantly with 7 scales of the AIMS, SIP, subjective health score, rating scale and standard gamble utilities, HAQ-S, pain (VAS), stiffness, functional index, physician's assessment of disease activity (Table 2). It should be noted that the significance level has not been adjusted for the number of comparisons made.

All statistical significant correlations occurred in "correct" directions, i.e., improvements in PET scores were associated with improvements in other outcomes.

Table 2. Spearman rank correlation coefficients between PET scores, and selected outcomes for 119 patients with AS and 70 patients with FMS

	Ankylosing spondylitis PET score		***Fibromyalgia*** PET score	
	Baseline	***Change***	***Baseline***	***Change***
AIMS scales:				
Mobility	0.002	0.12	0.25*	0.22*
Physical activity	0.22*	0.29*	0.40***	0.26*
Dexterity	0.14	-0.01	0.16	0.05
Social role	0.24*	0.07	0.38***	0.18
Social activities	-0.01	0.07	0.14	-0.10
Activities daily living	0.14	0.14	0.29*	-0.07
Pain	0.48***	0.41***	0.47***	0.36**
Depression	0.48***	0.12	0.34***	0.34**
Anxiety	0.49***	0.07	0.32***	0.24*
Health perception	0.41***	0.03	0.32***	0.10
Arthritis impact or patient's global	-0.38***	-0.21	-0.27*	-0.28*
SIP	0.43***	0.11	0.39***	0.11
HAQ	0.26*	0.11	0.38***	0.14
Pain (VAS)	0.44***	0.27*	0.37***	0.23*
Stiffness	0.44***	0.19	0.23*	0.31**
PET: Subjective health score	-0.41***	-0.36***	-0.47***	-0.51***
Utilities#:				
Rating scale	-0.36***	-0.28*	-0.39***	-0.44***
Standard gamble	-0.22*	0.14	-0.15	-0.25*
Functional index	0.30*	0.34***		
Articular index	0.10	0.12		
Enthesis index	0.12	0.07		
Chest expansion	-0.07	-0.16		
Flexion/extension	-0.13	-0.15		
Cervical rotation	-0.07	-0.29*		
Physical fitness	-0.17	0.14		
Physician assessed disease activity	0.29*	--.--		
Patient assessed improvement (VAS)	--.--	-0.13		
Patient's global health (NRS)			-0.52***	-0.32**

* $p<0.05$ ** $p<0.01$ *** $p<0.005$ # based on 59 AS and 73 FMS patients

In multiple regression analysis pain (29%), depression (11%) and arthritis impact scale (4%) of the AIMS together with stiffness (3%) contributed significantly in explaining variance of PET scores (Table 3).

Table 3. Ankylosing spondylitis: forward selection regression analysis with PET score as dependent variable and other baseline assessments as independent variables

Step	Variable entered	Partial R^2	F	p value
Dependent variable: PET score				
1	pain scale of AIMS	0.29	32.68	0.0001
2	depression scale of AIMS	0.11	14.10	0.0003
3	arthritis impact scale of AIMS	0.04	4.84	0.03
4	stiffness	0.03	4.11	0.04

PET score = -6.15(3.06)+[5.57(1.88)*pain]+[6.36(1.63)*depression]
-[2.17(0.99)*arthritis impact]+[2.78(1.09)*stiffness]
( ) = standard error

- *Fibromyalgia patients*

PET scores correlated significantly with 9 scales of the AIMS, SIP, subjective health score, rating scale utilities, mHAQ, pain (VAS), stiffness, and patient's global health (NRS) (Table 2). In multiple regression analysis patient's global assessment (NRS) (39%), pain scale of AIMS (15%), and stiffness (6%) contributed significantly in explaining variance of PET scores. However, the latter occurred in an unexpected direction (Table 4). Together the 3 variables accounted for 60% of total variance in PET scores (Table 4).

Table 4. Fibromyalgia: forward selection regression analysis with PET score as dependent variable and other baseline assessments as independent variables

Step	Variable entered	Partial R^2	F	p value
Dependent variable: PET score				
1	patient's global assessment (NRS)	0.39	29.70	0.0001
2	pain scale of AIMS	0.15	14.04	0.0005
3	stiffness	0.06	6.43	0.02#

PET score = 23.04(5.58)+[2.80(0.59)*pain]-[0.29(0.06)*global]-[1.39(0.55)*stiffness]
( ) = standard error # unexpected direction

Sensitivity to change of the PET questionnaire

- *Ankylosing spondylitis patients*

Changes in PET score correlated significantly with changes in 2 scales of the AIMS, and with changes in subjective health score, rating scale utilities, pain (VAS), functional index and cervical rotation (Table 2).

Multiple regression analysis with changes in PET scores as dependent variable and changes in other outcomes, subjective health score and the intervention as independent variables showed that 6% of total variance was explained by changes in pain scale of the AIMS (Table 5).

Table 5. Ankylosing spondylitis: forward selection regression analysis with change in PET score as dependent variable and changes in other assessments as independent variables

Step	Variable entered	Partial R^2	F	p value
Dependent variable: Change in PET score				
1	pain scale of AIMS	0.06	4.33	0.04

Change PET score = 1.40(0.76) + [0.56(0.28)*pain scale]
( ) = standard error

Calculation of efficiency of outcomes showed that physical fitness, thoracolumbar flexion and extension, and patient assessed improvement were the most sensitive measures, with $E > 0.7$ (Table 6). Cervical rotation and PET score were less sensitive with $0.6 < E < 0.7$. The efficiency of chest expansion, functional index and enthesis index was less with E values between 0.3 and 0.5, whereas all other outcomes were the least sensitive to change with $E < 0.3$ (Table 6).

- *Fibromyalgia patients*

Changes in PET score correlated significantly with changes in 6 scales of the AIMS, and with changes in subjective health score, rating scale and standard gamble utilities, pain (VAS), stiffness, and patient's global health (NRS) (Table 2).

Multiple regression analysis with changes in PET score as dependent variable and changes in other outcomes, subjective health score and the intervention as independent variables showed that 22% of total variance was explained by changes in the pain scale of the AIMS and 13% was explained by changes in subjective health score (Table 7).

Efficiency of outcomes showed that patient's global health was the most sensitive, with $E = 0.48$, followed by the social activities scale and social role scale of the AIMS ($E = 0.44$ and 0.35 respectively) (Table 6). The PET score and all other outcomes were less sensitive to change with $E < 0.3$ (Table 6).

Table 6. Mean changes* of PET scores and other outcomes

	Ankylosing spondylitis			*Fibromyalgia*		
	Mean	**SD**	**E$**	**Mean**	**SD**	**E$**
PET:						
PET score	-3.57	6.0	0.60	-0.65	7.4	0.09
Subjective health score	0.22	1.2	0.18	-0.40	1.4	0.29
Other outcome measures:						
AIMS scales:						
Mobility	0.02	0.42	0.05	0.22	1.42	0.15
Physical activity	-0.37	1.98	0.19	0.03	2.33	0.01
Dexterity	-0.03	1.02	0.03	-0.08	2.56	0.03
Social role	-0.09	0.56	0.16	0.23	0.65	0.35
Social activities	-0.28	1.19	0.24	-0.51	1.16	0.44
Activities of daily living	-0.03	0.34	0.09	0.05	0.72	0.07
Pain	-0.23	1.52	0.15	-0.30	1.52	0.20
Depression	-0.09	0.95	0.09	-0.29	1.55	0.19
Anxiety	-0.06	1.25	0.05	-0.37	1.45	0.26
Health perception	-0.17	1.45	0.12	-0.26	1.60	0.16
Arthritis impact	0.25	1.60	0.16	0.15	2.38	0.06
SIP	-0.63	3.33	0.19	-1.36	6.55	0.21
mHAQ	-0.01	0.19	0.05	0.07	0.37	0.19
Pain (VAS)	-4.53	23.38	0.19	-2.75	17.3	0.16
Stiffness	-0.66	18.40	0.04	12.3	46.9	0.26
Utilities#:						
Rating Scale	1.00	11.00	0.09	5.50	20.3	0.27
Standard Gamble	0.03	0.12	0.25	0.02	0.25	0.08
Functional index	-1.38	3.87	0.36			
Articular index	0.00	3.33				
Enthesis index	-0.79	2.56	0.31			
Chest expansion	0.50	1.16	0.43			
Flexion/extension	0.70	0.92	0.76			
Cervical rotation	13.78	19.94	0.69			
Physical fitness	-154.34	44.16	3.50			
Patient assessed improvement (VAS)	10.46	13.69	0.76			
Patient's global health (NRS)				0.85	1.76	0.48

* (followup - baseline)

\# based on 59 AS and 73 FMS patients

$ Efficiency: $E=d/SD_d$, where d is the mean change in the measure and SD_d is the standard deviation of the change measure[22]

Table 7. Fibromyalgia: forward selection regression analysis with change in PET score as dependent variable and changes in other assessments as independent variables

Step	Variable entered	Partial R^2	F	p value
Dependent variable: Change in PET score				
1	pain scale of AIMS	0.22	11.90	0.001
2	subjective health score (PET)	0.13	7.86	0.008

Change PET score = 2.76(1.15) + [1.21(0.37)*change pain]-[1.48(0.53)*change subjective health score]
( ) = standard error

DISCUSSION

The *feasibility* of applying the PET questionnaire was satisfactory in both patient groups. Many patients appreciated that the interview focussed to problems of importance to them. Despite efforts to help patients identifying problems, 10% of the patients with AS were not able to mention any problem at baseline, because they did not experience any limitation due to their disease. This was not the case in patients with FMS. This finding is of some concern in applying the PET questionnaire to patients with AS, because it will reduce the applicability of the PET questionnaire as an instrument in clinical studies or clinical practice.

In our study the PET questionnaire might be biased towards improvement as the number of problems was similar over time. When more problems would have been reported by the patient at followup this deterioration could not be captured in the PET. It has already been suggested to allow new problems to be reported at followup, although baseline values are of course missing for newly arising problems.[8]

The *construct validity* of the PET was satisfactory: pain scale of the AIMS contributed considerably in PET scores in patients with AS or FMS. Multiple regression analyses showed that variance in PET scores was well explained (47% in AS and 60% in FMS) by "other" outcomes, indicating that the PET also incorporates other aspects of health than represented in more traditional outcomes.

Sensitivity to change. In rheumatoid arthritis patients the PET questionnaire turned out to be more responsive than many traditional measures.[8] Recently we noted that in a trial on physiotherapy, patients with AS showed statistically significant improvements for spinal mobility (flexion/extension), physical fitness, and patient's assessed improvement. These 3 outcome measures now shown to be the most sensitive as assessed by efficiency analysis (Table 6). The efficiency of the PET score was less sensitive, however still reasonable with E = 0.6. In patients with FMS the efficiency of the PET score was very low (0.09) indicating the PET score to be insensitive in these patients. However, all other measures were insensitive in the patients with FMS.

In patients with AS the PET questionnaire appeared to be much more sensitive in a research setting than self-administered questionnaires, such as SIP, AIMS and HAQ-S. The PET questionnaire however, takes 20 minutes to apply. In clinical care the PET questionnaire is of value because it focusses on problems judged important by the patient on which therapy can be based.
In conclusion, obtaining patient priorities by PET questionnaire is feasible in patients with AS or FMS. Pain contributed considerably in explaining variance of PET scores. Construct validity testing of the PET questionnaire appeared satisfactory. Sensitivity to change was reasonable in patients with AS and low in patients with FMS. It is not yet known whether this is more related to the instrument than to this rheumatic condition. Further applications of PET-like questionnaires which assess patient priorities seems indicated in a variety of rheumatic diseases.

Acknowledgments
We thank all patients for their cooperation, Rob de Bie, Cootje Braakman, Carla Jonker, Annemiek Mackaay for their help interviewing the patients and Rachelle Buchbinder and Claire Bombardier for their help with the PET questionnaire.

REFERENCES

1 Bell MJ, Bombardier C, Tugwell P: Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.

2 Tugwell P, Bombardier C, Buchanan W, Goldsmith C, Grace E, Hanna B: The MACTAR patient preference disability questionnaire - An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987;14:446-51.

3 Hidding A, van der Linden S, Boers M, et al: Is group physical therapy superior to individualized therapy in Ankylosing Spondylitis? A randomized controlled trial. Arthritis Care Res 1993;6:117-25.

4 Santen van-Hoeufft M, Bolwijn P, Kleijnen J, et al: Is low impact fitness beneficial in fibromyalgia patients. Results of 2 randomized controlled trials. (in preparation)

5 Linden van der S, Valkenburg HA, Cats A: Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.

6 Wolfe F, Smythe HA, Yunus MB, et al: The American College of Rheumatology 1990. Criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. Arthritis Rheum 1990;33:160-72.

7 Bakker C, Rutten M, van Santen-Hoeufft M, Bolwijn P, van Doorslaer E, van der Linden S: Patient utilities in fibromyalgia and the association with other outcome measures. (submitted)

8 Buchbinder R and Bombardier C. Personal communication.

9 Guillemin F, Bombardier C, Beaton D: Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417-32.

10 Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.

11 Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ: Evaluation of the Dutch Arthritis Impact Measurement Scales (Dutch-AIMS) in patients with rheumatoid arthritis. Br J Rheumatol 1989;28:487-91.

12 Daltroy LH, Larson MG, Roberts WN, Liang MH: A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. J Rheumatol 1990;17:946-50.

13 Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP: Assessment of patient satisfaction in activities of daily living using a modification of the Stanford health assessment questionnaire. Arthritis Rheum 1983;26:1346-53.

14 Bakker C, Rutten-van Mölken M, Hidding A, van Doorslaer E, Bennett K, van der Linden S: Patient utilities in ankylosing spondylitis and the association with other outcome measures. J Rheumatol 1994;21:1298-1304.

15 Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B: Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988;15:302-7.

16 Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC: Studies with an enthesis index as a method of clinical assessment in ankylosing spondylitis. Ann Rheum Dis 1987;46:197-202.

17 Miller MH, Lee P, Smythe HA, Goldsmith H: Measurement of spinal mobility in the sagittal plane: New skin contraction technique compared with established methods. J Rheumatol 1984;11:507-11.

18 Tugwell P, Bombardier C: A methodological framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982;9:758-62.

19 Fleiss JL: The design and analysis of clinical experiments. New York: Wiley, 1986.

20 Pocock SJ: Clinical trials. A practical approach. New York: Wiley, 1983.

21 SAS/STAT Guide for personal computers version 6. SAS Institute Inc., Cary, NC, USA, 1985.

22 Anderson JJ, Chernoff MC: Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol 1993;20:535-7.

9

COST EFFECTIVENESS OF GROUP PHYSICAL THERAPY COMPARED TO INDIVIDUALIZED THERAPY FOR ANKYLOSING SPONDYLITIS. A RANDOMIZED CONTROLLED TRIAL

Carla Bakker, Alita Hidding, Sjef van der Linden, and Eddy van Doorslaer[1]

Department of Internal Medicine, Division of Rheumatology, University of Limburg, Maastricht; and Institute for Medical Technology Assessment[1], Erasmus University, Rotterdam, The Netherlands

Reprinted with permission of the Journal of Rheumatology 1994;21:264-8

COST EFFECTIVENESS OF GROUP PHYSICAL THERAPY COMPARED TO INDIVIDUALIZED THERAPY FOR ANKYLOSING SPONDYLITIS. A RANDOMIZED CONTROLLED TRIAL

ABSTRACT

Objective. Cost effectiveness analysis is helpful in setting priorities for funding of health care programs. We studied the cost effectiveness of supervised group physical therapy compared to unsupervised exercises at home in patients with Ankylosing Spondylitis (AS).

Methods. A total of 144 patients with AS (modified New York criteria; mean age: 43 years) were randomized to unsupervised daily individualized exercises at home for 9 months or the same plus supervised group physical therapy (3 h weekly). At baseline and after 9 months we measured spinal mobility (thoracolumbar flexion and extension), fitness (maximum work capacity by ergometry), and patient's global assessment of change as measured on a visual analogue scale. We used a questionnaire at baseline and a diary during the trial to measure AS related direct medical costs, such as doctor visits, paramedical treatment, medication and hospitalization.

Results. The mean effects of group therapy and home exercises were, respectively, +0.9 cm (16%) and +0.5 cm (9%) for mobility, +7 watts (4%) and -2 watts (-1%) for fitness, and +1.7 (34%) and +0.3 (6%) for global health. These 3 differences were significant ($p<0.01$ for mobility, $p=0.05$ for fitness and $p<0.01$ for global health). During the trial total medical costs decreased by an average of US $379 (44%) dollars for group therapy, and by $257 (35%) per patient per year for the "home" group. Additional costs of group therapy were estimated at $531 per patient per year ($177 for accommodation, $256 for therapist and $98 for materials). After the study 75% of the patients wanted to continue group physical therapy and were willing to pay for it.

Conclusion. Compared to therapy at home, additional benefits of group therapy cost $531 per year, but reduced direct medical costs by $122 per year. Hence, the beneficial effects of group therapy cost $409 per patient with AS per year.

INTRODUCTION

Health care costs are a growing concern to patients, physicians, and policy makers.[1-4] In making therapeutic choices medical as well as economic consequences have to be taken into account. This also applies to chronic diseases such as Ankylosing Spondylitis (AS) whose impact lies in impairing the quality rather than reducing the length of life.[5] AS is a chronic systemic inflammatory disorder of unknown etiology, affecting mainly the axial skeleton.[6]
There is currently no cure for AS, but most patients can be adequately managed. The aim of treatment of AS is to maintain or improve general functioning and quality of life. Nonsteroidal antiinflammatory drugs can reduce pain and inflammation, while regular exercises and physical therapy can improve mobility, strength and fitness.[7,8] A recent randomized controlled trial in AS showed supervised group physical therapy superior to unsupervised individualized exercises at home in improving thoracolumbar mobility, fitness, and patient's global assessment of change as measured on visual analogue scale.[9] During this 9-month trial, costs were registered in order to assess the financial implications of these therapeutic interventions. We report the cost effectiveness of group physical therapy to patients with AS compared with individualized physical therapy at home.

METHODS

A total of 333 patients with AS from 2 outpatient rheumatology clinics were asked to participate in the study. Of these, 163 (49%) gave written informed consent. All participants were examined by one rheumatologist to check for the inclusion criteria. We included patients fulfilling the modified New York criteria[10], with one or more of the following features: continuous symptoms of pain, stiffness, or functional limitations within the last 3 months, age below 70 and living within 25 km of a location of group physical therapy. Patients unable to engage in physical therapy, those with total hip replacement, pregnant patients, and those with severe hypertension [diastolic blood pressure > 100 mm Hg at rest], cardiovascular disease [history of ischemic event, angina pectoris, heart failure], severe lung disease, diabetes mellitus, renal failure, chronic liver disease, malignancy, recent major surgery, mental retardation or serious emotional disorders were excluded. Altogether 10 patients were excluded: 2 patients did not satisfy the modified New York criteria, 1 had a total hip replacement, 4 had cardiac problems, and 3 had emotional disorders, while an additional 9 patients stated that they were unable to exercise daily. Thus, 144 patients were available for the study.

Design and treatments
During 6 weeks before the study all patients received 12 sessions of supervised individualized physical therapy.[11] Afterwards, patients were randomized into 2 groups:
1) unsupervised daily individualized exercises at home; or 2) the same plus weekly group physical therapy.[9] The therapists encouraged the patients to continue the exercises at home for 30 min daily over the entire study period of 9 months.

Each weekly group therapy session consisted of 1 h of physical training, followed by 1 h of sporting activities and 1 h of hydrotherapy. The physical training included exercises to improve the mobility of the spine, hips, shoulders and peripheral joints and to strengthen the muscles of the trunk and legs.[8,9] During the sporting activities the therapists emphasized stretching of the back, for instance through volleyball or badminton. Hydrotherapy was given in heated water (mean: 31; range: 29 - 32°C) to reduce pain and to improve mobility of the spine, hip, shoulders and peripheral joints.
All patients continued to receive their usual reumatological care and medication during the study. The research team did not try to keep constant the dosis of any analgesics or antiinflammatory drugs used by the patient. However, the use of these drugs was monitored in order to detect beneficial effects.

Assessments of physical therapy
Primary predefined endpoints of therapy were spinal mobility, physical fitness, functioning and patient's global health.

- Spinal mobility was assessed using the 10 cm segment method to measure thoracolumbar flexion and extension (flex/ext).[12]
- Physical fitness or aerobic power was measured using an electronically braked bicycle ergometer (Jaeger ER 800, Breda, The Netherlands). During the test, heart rate was measured continuously using a sports tester (Support PE3000, Almere, The Netherlands). An incremental exercise test was used. The protocol started at 50 watts for 5 min and then increased by 10 watts every min. All subjects performed up to their subjective maximum workload.[13]
- Patient's global health was assessed by asking the patient to describe his or her perceived change in overall daily functioning after the 9-month treatment period, applying a 10 cm horizontal visual analogue scale (-5 = maximum worsening, 0 = no change, +5 = maximum improvement).[9]

The spinal mobility and fitness tests were assessed by one trained and "blinded" observer. These tests took place at the same time of the day for each patient. The 48 h test-retest reliability of the mobility and fitness tests was assessed beforehand in 19 randomly chosen patients. The intraclass correlation coefficients for test-retest reliability were high for mobility (0.96), and reasonable for fitness (0.72).[11]

Assessments of costs
A questionnaire at baseline and a diary during the trial were used to measure AS related direct medical costs. The questionnaire at baseline asked for costs incurred during the 1-year pretrial period. A diary was used to assess these costs during the 9-month experimental period. These costs were converted to annual figures. Both instruments aimed to measure 4 AS related direct medical costs (expressed in US dollars):

- Outpatient AS related visits to family physicians, rheumatologist, ophthalmologist, orthopedist and other visits (such as surgeon or dermatologist). Costs for family physicians were set at $15.90 per visit.[14] Fees for specialists used in this study represent a weighted average of the fees of the state health insurance system and the private insurance fees. All charges are all-out fees of continuation consults. This calculation was based on a telephone survey (SIG Health Information Utrecht, The Netherlands). Costs for rheumatologists were estimated at $15.05, those for ophthalmologists at $12.61, and for orthopedists at $14.05.

- Other care, comprising physical therapy, mental health care, and alternative care, including acupuncture, sauna and massage. Costs of physical therapy, using the same weighted average as for the costs for specialists, were assessed at $13.94. Costs of mental health care were put at $68.90, the normal fee of a psychiatrist per visit. Costs of alternative care were calculated from an average charge for the 3 different modes of treatment as determined by a telephone survey ($19.88).
- Purchases of medication, comprising nonsteroidal antiinflammatory drugs (NSAID), disease modifying antirheumatic drugs (DMARD, salazopyrine), analgesics, stomach medication, and eye medication. In the analysis, use of medication was converted into standard daily doses for each of these 5 groups. The costs of medicine consumption are based on current average retail prices.[15] The average costs for one day doses for each of the 5 groups of medication are $0.80 for NSAID, $0.77 for DMARD, $0.18 for analgesics, $1.53 for stomach medication and $3.13 for eye medication. The costs were added, yielding a total amount for medicine consumption.
- Hospitalization. Disease related days of admission for AS. Costs were calculated using the average of the all-out charges of the 3 participating rheumatology hospital departments ($324.40).

  After the study all patients who had received group physical therapy were asked if they wanted to continue this treatment and if so, whether they were willing to pay for it.

Statistical analysis

Data were summarized by mean, standard deviation (SD) or standard error of the mean (SEM). Baseline differences between groups regarding scores on the pretrial questionnaire were studied with Mann-Whitney test. At 9 months the groups were compared for mean improvement by t-test of change scores. Scores on the pretrial questionnaire and the diary were compared by Mann-Whitney test. The cost effectiveness was calculated using the formula given in Figure 1.

We predefined before the study spinal mobility, fitness, function and global assessment as primary outcomes. We defined an effect size of 20% for spinal flexion and extension or fitness as clinically important effects. We calculated that a sample size of 80 for each group would be required to provide a power of 90% to detect such an effect with an α probability of 5%.

Figure 1. Cost effectiveness formula

$$\frac{\Delta C}{\Delta E} \rightarrow \frac{C_{Int} - C_{Con}}{E_{Int} - E_{Con}}$$

ΔC = Costs of group therapy +
$((C_{Gr}\text{trial} - C_{Gr}\text{pretrial}) - (C_{Ind}\text{trial} - C_{Ind}\text{pretrial}))$
= \$531 + (-\$379 - (-\$257)) = \$409

ΔE = (flex/ext$_{Gr}$ 9 months - flex/ext$_{Gr}$ baseline) -
(flex/ext$_{Ind}$ 9 months - flex/ext$_{Ind}$ baseline) =
(6.2 - 5.3) - (5.8 - 5.3) = 0.4 cm
and
(fitness$_{Gr}$ 9 months - fitness$_{Gr}$ baseline) -
(fitness$_{Ind}$ 9 months - fitness$_{Ind}$ baseline) =
(177 -170) - (172 - 174) = 9 watt
and
change global health$_{Gr}$ 9 months -
change global health$_{Ind}$ 9 months) =
1.7 - 0.3 = 1.4

C = costs; E = effects; Gr = group physical therapy;
Int = intervention; Con = control; Ind = individualized exercises at home.

RESULTS

Description of study population

At baseline, 131 (91%) patients completed the pretrial questionnaire. During the 9-month experimental period 9 patients (8 in individual and 1 in group therapy) dropped out for the following reasons: moved (1), pregnant (1), spinal surgery (1), cardiac or lung disease (2), inability to exercise individually (4). After the 9-month experimental period 111 patients (77%) returned a completed diary. The cost effectiveness analysis is therefore based on 111 subjects (Table 1).

Table 1. Characteristics of 111 participants at baseline

Age (years) Mean (SD)	42.7(10.3)
Duration of disease (years) Mean (SD)	7.1 (7.2)
Patient's global assessment[*] Mean (SD)	7.2 (1.7)
Physician's global assessment[**] Mean (SD)	3.4 (0.8)
Socio-demographic characteristics (%)	
male	77
married	72
employed	73
income[#]:	
high	22
middle	56
low	22
education level[##]:	
high	32
middle	46
low	22

[*] 10-cm visual analogue scale (0 = worst; 10 = best)
[**] 5-point scale (1 = worst; 5 = best).
[#] US dollars net per month;
high: > 2000 $; middle: 1000 - 2000 $; low: < 1000 $
[##] Years of education (including primary school);
high: > 15 years; middle: 10 - 15 years; low: < 10 years

Baseline differences on costs between the treatment groups were not significant (Table 2).

Costs of supervised group physical therapy

In calculating the costs of supervised group physical therapy the preferred size of the group was defined as 12 patients, and the number of sessions at 40 a year. Costs for accomodation included weekly rent of a gymnasium (1 h physical training and 1 h sports, total costs US $1060) and 1 h weekly rent of a swimming pool with water at 32^0 degrees Celsius ($1060). Annual depreciation costs for materials included $636 for exercise mats, $382 for exercise balls and $159 for volleyballs. Costs of the physical therapist comprised $3074 for wages. Calculated costs per patient per year were $177 for accommodation, $98 for materials, and $256 for therapist. Hence, total additional costs of supervised group physical therapy were estimated at $531 per patient per year, or an average of $44 per month.

Table 2. Annualized disease related medical expenses in 111 AS patients

	Pretrial			Trial	
	Ind	Ind+gr	p*	Ind	Ind+gr
DOCTOR VISITS**					
Family physician	0.71±0.22	1.93±0.90	0.19	1.41±0.45	1.33±0.51
Rheumatologist	2.12±0.31	2.23±0.23	0.32	1.11±0.21	1.63±0.32
Ophthalmologist	0.14±0.12	0.30±0.30	0.46	0.56±0.11	0.72±0.14
Orthopedist	0.02±0.02	0.00±0.00	0.27	0.28±0.05	0.00±0.00
Other spec.	0.10±0.06	0.32±0.13	0.28	0.83±0.60	0.78±0.16
OTHER CARE**					
Physiotherapy	12.85±3.88	13.97±3.09	0.30	4.22±1.53	4.88±1.67
Mental Health	0.00±0.00	0.03±0.02	0.20	0.83±0.30	0.00±0.00
Alternative	1.28±0.83	1.62±1.40	0.33	4.66±1.69	2.63±0.90
MEDICATION#					
NSAID	0.49±0.07	0.60±0.07	0.27	0.57±0.09	0.59±0.07
DMARD	0.02±0.02	0.06±0.03	0.35	0.00±0.00	0.03±0.02
Analgesics	0.02±0.02	0.00±0.00	0.27	0.02±0.02	0.00±0.00
Stomach	0.05±0.03	0.03±0.02	0.85	0.03±0.03	0.01±0.01
Eye	0.04±0.03	0.07±0.03	0.56	0.02±0.02	0.02±0.02
HOSPITALIZATION##					
Admission	0.84±0.84	0.84±0.59	0.71	0.03±0.03	0.27±0.17

Values are mean ± SEM per patient per year.
* Mann-Whitney test for pretrial difference between individualized and group therapy.
** Mean number of visits or sessions.
\# Mean number of daily doses.
\#\# Mean number of admission days.
Ind = individualized exercises at home.
Gr = group physical therapy.

Effects of intervention

After 9 months the mean effects of group physical therapy and individualized exercises at home were, respectively, +0.9 cm (16%), and +0.5 cm (9%) for mobility, +7 watts (4%) and -2 watts (-1%) for fitness, and +1.7 cm (34%) and +0.3 cm (6%) for patient's global assessment of change in health.[9] These 3 differences were statistically significant ($p<0.01$ for mobility, $p=0.05$ for fitness, and $p<0.01$ for global assessment). Outcome measures of functional status, the Sickness Impact Profile[16] and the Health Assessment Questionnaire for the Spondyloarthropathies[17] did not show statistical significance[9] and were not used in the cost effectiveness analysis. Results of the significant measures indicated that, compared to individualized therapy at home, group physical therapy produced 7% extra increase in mobility, 5% extra increase in fitness, and 28% extra increase in global health.

Cost differences as effect of intervention

Total medical costs decreased by an average of $257 (35%) per patient per year for individualized exercises at home and by $379 (44%) for group therapy (Table 3). Most direct medical costs decreased more or increased less for those patients who had group therapy than for those who only exercised at home, with the exception of visits to rheumatologists and days of admission. The total decrease in medical costs was $122 more for those who had group

therapy than for those who exercised at home. The differences in decrease between the 2 groups were significant with regard to orthopedists, mental health treatments and total medical costs.

Cost effectiveness of intervention

Group physical therapy cost $531 per patient per year. The therapy reduced medical costs by $122. Therefore, 7% extra increase in mobility, 5% extra increase in fitness and 28% extra increase in global health cost $409 extra per patient per year (Figure 1). After the study 75% of the patients wanted to continue group physical therapy and were willing to pay for it.

Table 3. Annualized costs in 111 patients with AS

	Pretrial		Trial		Difference		
	Ind	**Ind+gr**	**Ind**	**Ind+gr**	**Ind**	**Ind+gr**	**p***
DOCTOR VISITS**							
Family physician	11	31	22	21	+11	-10	0.57
Rheumatologist	32	34	17	24	-15	-10	0.95
Ophthalmologist	2	4	7	9	+5	+5	0.84
Orthopedist	1	0	4	0	+3	0	<0.01
Other specialists	2	5	13	12	+11	+7	0.11
Total	48	74	63	66	+15	-8	0.54
OTHER CARE**							
Physiotherapy	179	195	59	68	-120	-127	0.49
Mental Health	0	2	57	0	+57	-2	<0.01
Alternative	25	32	93	52	+68	+20	0.42
Total	204	229	209	120	+5	-109	0.10
MEDICATION#							
NSAID	142	175	167	173	+25	-2	0.93
DMARD	6	15	0	10	-6	-5	0.85
Analgesics	1	0	2	0	+1	0	0.27
Stomach	26	16	15	2	-11	-14	0.42
Eye	46	75	23	19	-23	-56	0.41
Total	221	281	207	204	-14	-77	0.28
HOSPITALIZATION##							
Admission	271	271	8	86	-263	-185	0.70
TOTAL COSTS	744	855	487	476	-257	-379	0.05

Values are mean U.S. dollars per patient per year (1991 prices).
* Mann-Whitney test.
** Number of visits or sessions.
# Number of daily doses.
Number of admission days.
Ind = individualized exercises at home.
Gr = group physical therapy.

DISCUSSION

Our study compared the effects and costs of group physical therapy with individualized exercises at home. The decrease in the costs of physical therapy in our study was partly due to the exclusion by the study protocol of supervised individualized physical therapy for those randomized to exercises at home. Costs for alternative treatment during the trial increased especially for patients who had no group therapy.

In the pretrial period, hospitalization costs accounted for a reasonable part of total medical costs. Costs of doctor visits, paramedical treatment and medication in the pretrial period, $473 for exercises at home and $584 for group therapy, were increased by $271 hospitalization costs. In patients who exercised at home these hospitalization costs were caused by a total of 41 admission days for only 1 patient. In group therapy 2 patients together accounted for 51 days of admission. Days of admission were reduced during the trial to 1 day in individualized and to 16 days in group therapy. The mean costs and the mean difference in costs of hospitalization depended largely on a small number of expensive cases. This small number of patients accounted for the total decrease in medical costs in the patients exercising at home. The cost savings of group therapy, $8 for doctor visits, $109 for treatment and $77 for medication, were greatly outweighed by the difference in saved hospitalization costs.

Our study focussed on the direct medical costs, not because other effects are conceptually less relevant, but because our indirect costs assessments [e.g., decreased earnings or (un)paid help] were not very accurate.

In our analysis we fixed the price of exercises at home at zero. It is worthwhile to stress that supervised therapy beforehand is necessary to enable the patient to practice home exercises afterwards. Thus, treatment costs had already been made before the start of the trial.

In our analysis annualized medical costs before the trial were compared with the effects of a 9-month treatment period. A longer (1-year) treatment period might be more beneficial, as the costs of therapy remain the same.

The pretrial estimates of AS related direct medical costs were assessed and calculated retrospectively, while the expenses during the trial were collected prospectively. It is possible that the patient's recall of pretrial expenses may be less accurate than for expenses during the trial. However, this would probably only marginally influence the cost effectiveness analysis, because the analysis focussed on the differences in reduction of medical costs (Figure 1). However, it could mean that the reduction in medical costs would decrease and the costs of the beneficial effects of group therapy would be higher.

After this study we have to answer the question: is group physical therapy really cost-effective in AS? Doubilet, et al have pointed out that a treatment that is more expensive than the alternative can be considered cost-effective if it has an additional benefit worth the additional costs.[18] Indeed, cost effectiveness of a treatment should be related to the cost effectiveness of the *alternative* treatments. It might be argued that a 28% increase in global assessment is clinically relevant worth its price ($409). This might be supported by the fact that 75% of the patients wanted to continue group physical therapy and were willing to pay for it.

In a more complete cost effectiveness analysis of group physical therapy for a chronic disease like AS it is necessary to measure direct as well as indirect costs, e.g., decreased earnings.[19] Also, in a slowly progressive chronic disease a longer study period may be more appropriate to ascertain these longer term consequences. In addition, one might want to assess not only

disability related, but also more personality related effects, e.g., feelings of inadequacy or effect on health locus of control.[20]
Our study was done in The Netherlands. The results are applicable to countries with comparable health care systems and comparable health care costs. In countries with other health systems, where the current health care delivery system is much more costly, the beneficial effects of group therapy may cost considerable more than $409 per patient per year.
In summary, it can be said that compared to individualized therapy at home, group physical therapy produced an extra increase of 7% in spinal mobility, 5% in fitness and 28% in global health, at a cost of $531 per year, and decreased medical costs by $122 per year. The beneficial effects of group therapy in AS thus required extra outlays of $409 per patient per year.

ACKNOWLEDGEMENTS

We thank Jacqueline Mertens for examining the patients, Jan Thomassen for instructing group physical therapy, Xandra Gielen for preparation of the diaries, and Gerard Engel for helpful comments.

REFERENCES

1 Department of clinical epidemiology and biostatistics, McMaster University Health Science Centre: How to read clinical journals: VII. To understand an economic evaluation (part A). Can Med Assoc J 1984; 130: 1428-33

2 Department of clinical epidemiology and biostatistics, McMaster University Health Science Centre: How to read clinical journals: VII. To understand an economic evaluation (part B). Can Med Assoc J 1984; 130: 1542-49

3 Detsky AS, Naglie IG: A clinician's guide to cost effectiveness analysis. Ann Intern Med 1990; 113: 147-54

4 Udvarhelyi S, Colditz GA, Rai A, Epstein AM: Cost effectiveness and cost-benefit analyses in the medical literature. Are the methods being used correctly? Ann Intern Med 1992; 116: 238-44

5 Thompson MS, Read JL, Hutchings HC, Paterson M, Harris ED: The cost effectiveness of Auranofin: Results of a randomized clinical trial. J Rheumatol 1988; 15: 35-42

6 Khan MA, van der Linden S: Ankylosing Spondylitis and other Spondylarthropathies. Rheum Dis Clin North Am 1990; 16: 551-78

7 Kraag G, Stokes B, Groh J, Helewa A, Goldsmith C: The effects of comprehensive home physiotherapy and supervision on patients with Ankylosing Spondylitis. A randomized controlled trial. J Rheumatol 1990; 17: 228-33

8 Spring H: Funktionsorientierte Gymnastik und Sport bei der Spondylitis Ankylosans. In: Spring H, ed. Spondylitis Ankylosans. Bern: Verlang Hans Huber, 1989: 117-32.

9 Hidding A, van der Linden S, Boers M, et al: Is group physical therapy superior to individualized therapy in Ankylosing Spondylitis? A randomized controlled trial. Arthritis Care Res 1993; in press

10 van der Linden S, Valkenburg HA, Cats A: Evaluation of diagnostic criteria for Ankylosing Spondylitis: A proposal for modification of the New York criteria. Arthritis Rheum 1984; 27: 361-8

11 Hidding A, van der Linden S, de Witte L: Therapeutic effects of individual physical therapy in Ankylosing Spondylitis related to duration of disease. Clin Rheumatol 1993; in press

12 Miller MH, Lee P, Smythe HA, Goldsmith H: Measurement of spinal mobility in sagittal plane: New skin contraction technique compared with established methods. J Rheumatol 1984; 11: 507-11

13 Jones NL, Makrides L, Hitchcock C, Chypchar T, McCartney N: Normal standards for an incremental progressive cycle ergometer test. Am Rev Respir Dis 1985; 131: 700-8

14 Rutten FFH, van Ineveld BM, van Ommen R, van Hout BA, Huijsman R: Kostenberekening bij gezondheidszorgonderzoek; richtlijnen voor de praktijk. STG brochure. Utrecht: van Arkel, 1993

15 van der Kuy A, et al: Farmacotherapeutisch Kompas 1991: medische farmaceutische voorlichting. Amstelveen: Centrale Medisch Pharmaceutische Commissie Ziekenfondsraad, 1982

16 Bergner M, Bobbit RA, Kressel S, Pollard WE, Gilson BS, Morris JR: The Sickness Impact Profile: Conceptual formulation and methodology for the development of a health status measure. Int J Health Serv 1976; 6: 393-415

17 Daltroy HL, Larson MG, Roberts WN, Liang MH: A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. J Rheumatol 1990; 17: 946-50

18 Doubilet P, Weinstein MC, McNeil BJ: Use and misuse of the term "cost effective" in medicine. N Engl J Med 1986; 314: 253-6

19 Meenan RF, Yelin EH, Henke CJ, Curtis DL, Epstein WV: The costs of Rheumatoid Arthritis: A patient-oriented study of chronic disease costs. Arthritis Rheum 1978; 21: 827-33

20 Lawrence VA, Tugwell P, Gafni A, Kosuwon W, Spitzer WO: Acute low back pain and economics of therapy: The iterative loop approach. J Clin Epidemiol 1992; 45: 301-11

10

GENERAL DISCUSSION

GENERAL DISCUSSION

Physician assessed and patient reported measures can be classified as either specific or generic in each of 3 areas of focus: health status, disease and patient (See cover). Outcome measures are multidimensional and can be generic in one area and focussed in another. From our review of outcome measures used currently in ankylosing spondylitis, it was clear that some areas of the total spectrum of health status were relatively overrepresented, whereas other areas were not well covered (**chapter 2**). Especially self-assessed measurement of functional ability, patient priority assessment, economic evaluation, and drug toxicity were frequently lacking.
In this thesis we concentrated on the measurement of patient priorities. Such measures can have major influence on clinical decision making. Until now, the physician has implicitly decided to continue or stop treatment on the basis of clinical history, physical examination and supplementary investigations (laboratory, X-ray). We support the view that it is more appropriate to ask the patients themselves to balance positive effects (for instance less pain) and negative effects (for instance nausea) of treatment in one overall value. Who other than the patient is a better judge of whether the achieved improvement outweighs the negative effects? An utility value is a value in which the patients is asked to evaluate their overall health status between 0 (equal to death) and 1 (equal to perfect health). Utilities do not show the dimensions in which improvement or deterioration occurs. This valuable information can be obtained by the simultaneous use of more focussed measures. Utility measures are also particularly relevant if the economic implications of an intervention are a major focus of investigation.[1] Nowadays, health care providers are frequently asked to justify the limited resources devoted to treatment. Therefore, we felt it would be very relevant to concentrate on utility measurement (chapter 3 to 7).
The unique feature of utility measurement is that it is generic in the health and disease area. In the patient area it is focussed when one assesses utilities of individual patients and generic when eliciting utilities from the general population. The utility value allows for comparison across different diseases and interventions, if the same dimensions of health and the same measurement techniques are used to obtain utilities.

Chapter 3 to 7 describe utility measurement by means of the Maastricht Utility Measurement Questionnaire (MUMQ) in patients with ankylosing spondylitis or fibromyalgia. The MUMQ can be explained in 3 steps: 1) definition of health, 2) description of health states, and 3) valuation of health states.

• Definition of health

Health is a broad concept and can be defined in many dimensions. In the MUMQ, health has been defined by 6 dimensions: physical state and mobility, self-care, emotions, leisure activities, pain, and side effects of treatment. It is important that the dimensions used to describe health are generic enough to cover all (rheumatic) diseases. On the other hand it is important that the dimensions are relevant to the particular disease under study. The latter is of major importance when using the utility measurement in the evaluation of an intervention. This also means that one would like to use measures that are sensitive to change.

A translation and adaptation of the original McMaster dimensions[2] was used in the studies described in chapter 4 and 6. In a later study described in chapter 5, however, we felt it necessary to modify these original dimensions to a more generalizable setting or description. The dimensions used in MUMQ are appropriate for rheumatic diseases. Pain is very prominent in rheumatic diseases and therefore covered in a separate dimension. By contrast, in chronic obstructive pulmonary diseases, for example, pain is of less prominent and in that case a separate dimension on dyspnoea is more appropriate. Moreover, no dimensions on sleep disturbance, and psychological or emotional distress were yet included in the MUMQ. We now suggest that adding a dimension on sleep disturbance, and psychological distress or emotional well being will further improve face validity of the health status instruments used in rheumatic patients.

• Description of health states

In the MUMQ, health states were described by combining 6 levels of severity, one of each dimension. Besides death and perfect health, a mild, moderate, and severe marker state were created, which describe a mild, moderate and severe state of ankylosing spondylitis or fibromyalgia. These marker states are valuable during the measurement process, as they encourage the patient to consider a broad range of possibilities before determining their own health on the spectrum of possibilities.[3] The marker states are also called reference states. However, the marker states can be confusing for the patient as they describe abstract hypothetical states of illness that they have never experienced or are unable to imagine. Moreover, it appeared difficult for the patient to integrate all 6 dimensions into one state of illness. For instance, some patients with fibromyalgia focussed mainly on the pain dimension. Because of this difficulty for patients to conceptualize marker states, it might be important in future research to restrict the use of marker states to a maximum number of one or two. Patients found it also difficult to imagine the severe marker state and to distinguish it from death. This resulted in about 50% of the patients valuing the severe marker state at least once (at baseline and/or at followup) worse than death. A number of patients changed their opinions as to whether the severe state was better or worse than death (chapter 7). Possibly, the health state of a severely ill patient is so hard to imagine that patients put all their effort into understanding it, instead of paying attention to what was really being asked. When measuring how much worse the severe marker state was than death, highly negative utilities occurred that greatly influenced the mean utility (chapter 7). These findings and the difficulties of handling extreme negative utilities may argue in favor of using death instead of using the severe marker state as the worst outcome of the gamble.

Another advantage of using marker states is that it enables the measurement of test-retest reliability of the utility measurement instrument at the repeated followups, since the reference (marker) states are expected to remain the same throughout the study.[3] However, remarkable

changes in preferences occurred in fibromyalgia for the mild marker state after 3 months (chapter 7) and the severe marker state after 6 months (chapter 6), as well as changes in preferences in ankylosing spondylitis for the severe marker state after 3 months (chapter 7). These changes possibly result from changes in the patient's perception of his/her own health state that in turn induces the valuations for the reference (marker) states to change too. In such situations it can not be expected that the reference states remain stable over time (chapter 6).

- Valuation of health states

Utilities were assessed using the rating scale and standard gamble method. The standard gamble method may be seen as the gold standard in utility measurement as it is directly based on the Von Neumann-Morgenstern utility theory.[4] However, the rating scale method has been used far more often, probably because it is easier to apply and to understand. It should be noted that in the standard gamble method health states are valued under the assumption of risk, as opposed to the rating scale method were risk is not included in the measurement process.

In the MUMQ we used a rating scale that looked like a thermometer with 'perfect health' equal to 100 at the top and the 'severe marker state' equal to 0 at the bottom. The standard gamble method was performed with a probability wheel as a prop. A 2-step utility assessment was applied to avoid that rheumatic patients are directly confrontated with a risk of dying.

The *feasibility* of the rating scale and standard gamble in both ankylosing spondylitis and fibromyalgia patients was generally satisfactory (chapter 4). Test-retest *reliability* was on average somewhat lower for the rating scale than for the standard gamble method (chapter 4,6). However, this may have been influenced by recoding the negative standard gamble utilities for the severe marker state to zero. As described before, reliability of utility measurement was also assessed by the stability of marker states of disease over time (chapter 5,6), and by the measurement error (the standard error of measurement) (chapter 7). The relative stability of preferences for marker states over time and the considerable measurement errors, that indicate the relative stability of individual patient preferences, support the notion that utilities may be less useful for individual decision-making than for group decision-making.

Construct validity was tested by Spearman correlations of utilities with scores on other outcomes such as, Sickness Impact Profile, Arthritis Impact Measurement Scale, Health Assessment Questionnaire, pain, stiffness, global health, spinal mobility and fitness. Construct validity seemed to be higher for the rating scale than for the standard gamble method (chapter 5,6) in both patients with ankylosing spondylitis or fibromyalgia. Regression analyses showed that the patient's global assessment explained a fair 41-59% of rating scale utilities but only 10-11% of standard gamble utilities in ankylosing spondylitis and fibromyalgia (chapter 5,6). These findings support the view that rating scale utilities more closely resemble global assessment. The methodological advantage of standardized rating scale utility measurement over non-standardized global assessment is that utilities provide numerical values which allows patient outcomes of different diseases or resulting from various health care interventions to be compared across patients and diseases.

Sensitivity to change was tested by Spearman correlations of changes in utilities with changes of scores on other outcomes as mentioned earlier. Changes in rating scale utilities correlated to a higher degree with changes in other outcomes than changes in standard gamble utilities in both diseases.

Based on the results of construct validity and sensitivity to change as described in chapter 5 and 6, the validity of the standard gamble is lower than the validity of the rating scale. In our opinion the standard gamble utilities might largely reflect measurement error, because only 10-

11% of standard gamble utilities could be explained by other (accepted and validated) outcome measures. This is a very low value and it might mean that the standard gamble method does not measure health status, but something completely different. Therefore, we prefer to use the rating scale to obtain utilities: a thermometer with "perfect health" equal to 100 at the top and "death" equal to 0 at the bottom. The newly introduced EUROQOL is also based on a rating scale.[5] The EUROQOL uses 5 dimensions of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, with 3 items each. When using these dimensions for rheumatic conditions it might be important to distinguish pain from discomfort in a separate dimension. As stated before, pain is very dominant in rheumatic patients. In our opinion, it should be kept in a separate dimension. Discomfort could also comprise other complaints, like for instance stiffness. Also the anxiety dimension should be separated from depression. The use of an extra dimension on side effects would be appropriate in chronic rheumatic diseases. The 5 EUROQOL dimensions are probably not very sensitive to improvement, because more aspects of function are joined in 1 dimension. For example, the dimension "usual activities" comprises work, study, housework, family, or leisure activities. Moreover, each dimension consists only of 3 items, for example, no problems in walking about, some problems in walking about and confined to bed. The grading steps between these 3 items are very big. Specially in chronic (rheumatic) diseases small but clinically important changes might occur and they could be easily missed when using only 3 items and (too) few grading levels.
When applying the standard gamble method, we suggest to use a 1-step procedure, i.e., to value the patient's health status directly between perfect health and death. We suggest to do so despite the fact that death is usually not a realistic outcome for most chronic rheumatic patients. Indeed, the second step of the standard gamble resulted frequently in very confusing and inconsequent answers (chapter 7).

Chapter 8 describes the Problem Elicitation Technique (PET) questionnaire. The PET questionnaire is a preference disability questionnaire, i.e., it focusses on activities that are limited by the disease and considered important by the patient. Tugwell suggested that the use of focussed responsive instruments may reduce the sample size needed in clinical trials, because he believes these instruments have an increased potential compared to existing questionnaires for demonstrating changes in disability in clinical trials by focussing on those activities directly affected by the disease and judged important by the patient.[6] In patients with ankylosing spondylitis, the PET indeed showed to be reasonable responsive (chapter 8). By contrast, in patients with fibromyalgia the PET questionnaire showed no responsiveness at all (chapter 8). It would be of interest to know whether this low responsiveness is related to this particular rheumatic condition (fibromyalgia) or to the instrument itself. In the former situation it might be important to assess the sensitivity to change of the PET questionnaire in other rheumatic diseases. Moreover, 10% of the patients with ankylosing spondylitis was not able to mention any problem at baseline, because they did not experience any limitation due to their disease. By contrast, in patients with fibromyalgia this was not the case. Patients with fibromyalgia easily came up with a number of problems. In our opinion this might be rather characteristic for this condition. Again, in applying the PET questionnaire to 2 different rheumatic conditions, the responsiveness was influenced by the patients' capability of mentioning problems at baseline. This would also suggest further research of sensitivity to change of PET-like questionnaires in different rheumatic conditions.
The feasibility as well as the construct validity of the PET questionnaire was satisfactory both for ankylosing spondylitis patients and fibromyalgia patients. Interestingly, in both diseases pain

explained a considerable amount of variation in the PET scores (chapter 8).
To improve the convenience of the PET questionnaire for both patients and investigators, we suggest that the PET should concentrate on a limited number of problems. For instance, the 5 most important problems mentioned at baseline should be followed over time. Another concern in our opinion is the PET score. In our study this was calculated by multiplying the difficulty (or severity or frequency) score by the importance score. These results were summed up for all problems and divided by the number of problems for each patient. However, why should the importance score be given the same weight as the difficulty (or severity or frequency) score? It is assumed that when patients improve at followup, their difficulty score decreases and that their problem is of less importance and thus the importance score decreases too. It is quite confusing when patients feel better at followup and their difficulty score decreases, but they feel the problem is now of major importance than at baseline. The PET score could turn out to be higher at followup than at baseline, even when the patient feels better at followup. Moreover, in our study the problem indicated as the most important one could turn out to get the same weight (the same PET score) as the problem indicated as the less important, dependent from the difficulty (severity or frequency) score. For example, a problem which is given level 3 for difficulty and is considered as a most important problem (importance level 7) results in a PET score of 21. The same score results when a certain task is unable te perform (level 7), but is less important (importance level 3). Perhaps a scoring system as the one used in the Sickness Impact Profile questionnaire[7], in which different weights are used for the different dimensions, should also be developed for the PET questionnaire. Perhaps different problems and their importance should have different weights.

In **chapter 9** the cost-effectiveness of group physical therapy was compared to unsupervised exercises at home in patients with ankylosing spondylitis. Group physical therapy cost $531 per patient per year. Direct medical costs for patients receiving group physical therapy were reduced by $122 per patient per year. After 9 months, group physical therapy produced an extra improvement of 7% in spinal mobility, of 5% in fitness and of 28% in patient's assessment of global health. Hence, the beneficial effects of group physical therapy cost 409$ per ankylosing spondylitis patient on a yearly basis or about $ 10 per weekly therapy session.
According to Doubilet, the cost-effectiveness of a treatment should be related to the cost-effectiveness of alternative treatments.[8] Weekly supervised individualized therapy in 30 minute sessions is more expensive than 3 hours of weekly group physical therapy ($575 and $531 per 40 sessions per patient per year respectively). It might be stated that a 28% increase in global assessment is worth its price ($409). At the end of the study 75% of the patients wanted to continue group physical therapy, and were willing to pay for it. Unfortunately, we did not ask how much money they were willing to pay for it in relation to their income. In fact, this evaluation includes both direct and indirect costs. Indirect costs include decreased earnings by the patient.[9] The indirect costs are of major importance in chronic diseases and it is a shortcoming of this study that we did not include these costs. We actually attempted to measure them, but (unfortunately) the results could not be used, because the questions concerning the patient's indirect costs were asked differently at baseline in the questionnaire than in the diary. In future cost-effectiveness analyses both direct and indirect costs should be included.
Another shortcoming of our study is that we were not able to relate the cost-effectiveness results to cost-utility or QALY's. The utility assessment was focussed on a subgroup of 59 patients. This was necessary because at followup 2 (of in total 4) interviewers were not able to participate anymore and these interviews were done by the 2 remaining interviewers. This

resulted in an interviewer effect. Therefore, we restricted the utility analysis to the 59 patients who were seen by the same interviewer at both assessments.
This study was done in the Netherlands. The financial findings are therefore most relevant to countries with comparable health care systems in so far as the costs are concerned.
It must be stressed that in a slowly progressive chronic disease a longer study period than 9 months may be more appropriate to ascertain long term consequences. In addition, it might be desirable to value not only disability-related, but also more personality-related effects, e.g., feelings of inadequacy or positive effects on the patient's health locus of control.[10]

Patient-oriented outcome assessment is very important to assess the impact of chronic diseases on the patient. In this thesis some priority measures applied to patients with ankylosing spondylitis or fibromyalgia were discussed. The assessment of patient priorities/preferences by means of utilities and PET questionnaire gave rise to more questions. Therefore, in our opinion further testing of these measures in different rheumatic diseases is needed, before these can be applied fruitfully on a large scale in clinical practice or in health service research.

REFERENCES

1 Guyatt GH, Feeny DH, Patrick DL: Measuring health-related quality of life. Ann Intern Med 1993;118:622-9.

2 Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991;(suppl)12:118-28.

3 Torrance GW, Feeny D: Utilities and quality-adjusted life years. Intl J of Technology Assessment in Health Care 1989;5:559-75.

4 von Neumann J, Morgenstern O: Theory of games and economic behavior. Princeton, NJ: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).

5 The EuroQol group: A new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.

6 Tugwell P, Bombardier C, Buchanan W, Goldsmith C, Grace E, Hanna B: The MACTAR patient preference disability questionnaire - An individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. J Rheumatol 1987;14:446-51.

7 Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.

8 Doubilet P, Weinstein MC, McNeil BJ: Use and misuse of the term "cost-effective" in medicine. N Engl J Med 1986;314:253-6.

9 Meenan RF, Yelin EH, Henke CJ, Curtis DL Epstein WV: The costs of rheumatoid arthritis: A patient-oriented study of chronic disease costs. Arthritis Rheum 1978;21:827-33.

10 Lawrence VA, Tugwell P, Gafni A, Kosuwon W, Spitzer WO: Acute low back pain and economics of therapy: The iterative loop approach. J Clin Epidemiol 1992;45: 301-311.

11

SUMMARY

SUMMARY

This thesis concerns patient-oriented outcome assessment in rheumatic diseases.
Chapter 1 provides an introduction and guide to this thesis. Measuring outcome or health status is important for assessing the impact of chronic diseases on the patient. Comprehensive outcome or endpoint measuring should include all those components of health status that are important to the patient and physician and that are relevant to the intervention assessed.
Health status can be described in 5 dimensions: Death, Disability, Discomfort, Drug (or therapeutic toxicity), Dollar costs. The scope of health status can be either specifically oriented, focussing on only one dimension, or broadly oriented, focussing on several dimensions. Further, different diseases may deduct different aspects of health status, whereas individual patients may have different priorities regarding their health status. Physician-assessed and patient-reported measures can be classified as either specific or generic in each of 3 areas of focus: health status, disease and patient (See cover). A measure is specific in the health status area when it measures only one dimension, and generic when it measures several dimensions. A measure is specific in the disease area when it is applicable to only 1 disease (e.g., ankylosing spondylitis). It becomes less focussed when it is applied in a group of diseases (e.g., all arthritides, all cancers) and completely generic when applicable to all diseases. Similarly, measures that are specific in the patient area refer to single patients. Less specifically focussed measures refer to subgroups (e.g., the elderly) and generic measures refer to all possible patients. Obviously, outcome measures are multidimensional and can be generic in one area and focussed in another.

Chapter 2 reviews outcome measures used today in ankylosing spondylitis. Of the 43 studies reviewed, 79% contained physician assessed measures. Patient reported measures were mentioned in 67%. Most physician assessed measures (67%) focussed on disability (spinal mobility), most patient reported measures (65%) focussed on discomfort (pain and stiffness). Single item global assessment by physician or patient, the most generic measure, was reported in 16% and in 40% of the studies respectively. Other general measures of health status were only occasionally used. In particular, self-assessed measurement of functional ability was frequently lacking. Only 1 study reported a measure which specifically addressed the patient's priorities regarding treatment risks, and only 3 studies included an economic analysis. Drug toxicity reports missed informative detail. This review clearly showed that some areas of the total spectrum of health status were not well covered, whereas other areas were overrepresented.

The chapters 3 to 8 deal with the measurement of patient priorities.
Chapter 3 gives an introduction to utility assessment in rheumatology. The unique feature of utility measurement is that it is generic in the health and disease area of the cube figure (See cover). In the patient area it is focussed when measuring utilities of individual patients and generic when eliciting utilities from the general population. The utility allows for comparison across different diseases and interventions. This chapter describes 2 approaches to utility measurement. The first approach is to classify patients into categories based on their responses to questions about their functional status (for example, the Quality of Well being questionnaire and the Sickness Impact Profile). The second approach to utility measurement is to ask the patients to directly assign one value to their overall health. The 4 methods used most frequently (rating scale, standard gamble, time trade-off and willingness to pay) were described.

Chapter 4 describes the feasibility of utility assessment by rating scale and standard gamble method in patients with ankylosing spondylitis (AS) or fibromyalgia (FMS). In the context of 2 randomized controlled trials the Maastricht Utility Measurement Questionnaire (MUMQ) was applied by trained interviewers to 59 patients with AS at the 9 months' followup and 86 patients with FMS at baseline. The MUMQ is a Dutch translation and adaptation of the McMaster Utility Measurement Questionnaire developed by Bennett and Torrance and consists of a rating scale and standard gamble method. The feasibility of these methods in both patients with AS and FMS was generally satisfactory. All patients completed the interview. Four (2.8%) of all 143 patients gave inconsistent answers: one on the rating scale and 3 on the standard gamble. The duration of the baseline interview was about 9-12 minutes (SD 3.6-5.4) for the rating scale and about 12-14 minutes (SD 2.9-3.8) on the standard gamble. Four weeks test-retest reliability for the patient's own health state measured by intraclass correlation coefficients was 0.56 for the rating scale and 0.66 for the standard gamble technique. Patients with AS valued their personal health state on the rating scale (0-100) considerably higher than patients with FMS (AS:69 and FMS:54). Standard gamble utility values (0-1), however, were about the same at a higher level (AS:0.86 and FMS:0.83). These data supported the view that utility measurement is sensitive to the method chosen to elicit patient well being.

Chapter 5 and 6 compare utilities derived by rating scale and standard gamble method in patients with AS and FMS, relate these values to other outcome measures, and describe the sensitivity to change of utilities relative to changes in other outcomes.
In both patient groups the MUMQ was applied twice: first at baseline and again after 6 (FMS) or 9 (AS) months' followup. In AS the analysis was restricted to the 59 patients who were seen by the same interviewer at both assessments. In FMS a total of 73 patients was assessed by the same interviewer at both assessments.
Construct validity seemed to be higher for the rating scale than for the standard gamble method in both patients with AS or FMS. For the rating scale Spearman correlation coefficients ranged from 0.25 to 0.68 in AS and from 0.23 to 0.62 in FMS; for the standard gamble it ranged from 0.23 to 0.36 in both diseases. Regression analyses showed that patient's global assessment explained 41-59% of rating scale utilities and 10-11% of standard gamble utilities in patients with AS or FMS. These findings support the view that rating scale utilities more closely resemble global assessment. The methodological advantage of standardized rating scale utility measurement over non-standardized global assessment is that utilities provide numerical values which allows patient outcomes of different diseases or resulting from various health care interventions to be compared across patients and diseases.

In both diseases, changes in rating scale utilities correlated to a higher degree with changes in other outcomes than changes in standard gamble utilities.

Chapter 7 provides a detailed description of methodological issues of patient utility measurement in 2 randomized controlled trials involving 144 patients with AS and 85 patients with FMS. In both trials the MUMQ was applied at baseline and at 2 followup assessments. Patients were asked to value their own health state and a light, moderate and severe marker or reference state by means of the rating scale and standard gamble method.
It was confirmed that standard gamble scores are consistently higher than rating scale scores for both the patient's own health state and the marker states. The 3 months' test-retest reliability for the reference states measured by intraclass correlation coefficients ranged from 0.24 to 0.33 using the rating scale and 0.43 to 0.70 on the standard gamble. Although the reproducibility is not high, the mean scores are fairly stable over time. Mean standard gamble scores tend to differ depending on how the measurements are taken. Utilities elicited with 'chained gambles' were significantly higher than utilities elicited with 'basic reference gambles'. On the individual level some inconsistent responses occurred. However, more than 70% of them fell within the bounds of the measurement error which ranged from 0.11 to 0.13 on the standard gamble (0-1 scale) and from 8 to 10 on the rating scale (0-100 scale). The large number of negative utilities for the severe marker state, which was used as an anchor point in the chained gambles, and the magnitude of these negative utilities (down to -19) lead us to favor using death as the anchor point in the standard gamble.

Chapter 8 describes another way to assess patient priorities by means of the Problem Elicitation Technique (PET) questionnaire. This questionnaire deals only with activities that are directly limited by the disease and judged important by the patient. The PET questionnaire was applied by trained interviewers to 134 patients with AS and 73 patients with FMS at baseline and at 9 and 6 months' followup respectively. A PET score was calculated at each assessment.
Patients with FMS identified more problems (mean 6.8) than patients with AS (mean 4.4). Moreover, patients with AS could more often not identify any problem at baseline than patients with FMS (10% compared to 1%). The PET score improved from 14.9 to 11.3 ($p=0.0001$) in AS patients but did not change from 21.8 to 21.1 ($p=0.24$) in FMS patients. Construct validity testing of the PET score showed statistically significant ($p<0.05$) correlations with Arthritis Impact Measurement Scale (AIMS), utilities, Sickness Impact Profile (SIP), Health Assessment Questionnaire (HAQ), pain, stiffness, and patient's global health in both AS and FMS patients (r's varying from 0.22 to 0.66). By multiple regression pain explained 29% of the variance in PET scores among patients with AS. In FMS, patient's global assessment accounted for 39% of total variance of PET scores, whereas pain explained another 15%. Changes in PET scores correlated significantly ($p<0.05$) with changes in AIMS, utilities, pain, stiffness, and patient's global health in both patients with AS and FMS (r's varying from 0.22 to 0.51). Some 6% of the variance in changes in PET scores was explained by changes in pain in patients with AS and for 35% by changes in pain and subjective health in patients with FMS. Assessment of sensitivity to change revealed that the efficiency of the PET score was 0.6 in patients with AS and 0.09 in patients with FMS. Compared to other outcomes this was reasonable in patients with AS but low in patients with FMS.
It was generally feasible to obtain patient priorities. In both patients with AS or FMS construct validity of the PET questionnaire was satisfactory. The PET was much more sensitive to change in patients with AS than in patients with FMS.

Chapter 9 provides a comparison of the cost-effectiveness of supervised group therapy with that of exercises at home in 144 patients with AS. Compared to exercises at home, group therapy produces an extra increase of 7% in spinal mobility and of 5% in fitness, and an extra improvement of 28% in global health at a cost of $531 per year, while reducing direct medical costs by $122 per year. Hence, the beneficial effects of group therapy cost $409 per patient with AS on a yearly base or about $10 per weekly therapy session (40 sessions per year).

Chapter 10 provides a general discussion of the study described in this thesis.

12

SAMENVATTING

SAMENVATTING

Dit proefschrift behandelt patiënt-georiënteerde uitkomstmetingen bij reumatische aandoeningen. **Hoofdstuk 1** geeft een inleiding tot dit proefschrift. Het is van belang de gezondheidstoestand of de 'uitkomst' (outcome) te meten, om zo de impact van chronische aandoeningen op de patiënt vast te stellen. Om de uitkomst, of het eindpunt goed te kunnen bepalen moet het die aspecten van de gezondheidstoestand omvatten die belangrijk zijn voor de patiënt en de arts en tevens moeten ze relevant zijn voor de meting. Uitkomstmetingen kunnen worden ingedeeld als door de arts gemeten, door de patiënt gerapporteerd of 'overig' (bijvoorbeeld laboratorium waarden en röntgen foto's).
De gezondheidstoestand kan beschreven worden in vijf dimensies: dood, verminderd functioneren, klachten of symptomen zoals pijn en stijfheid, bijwerkingen van medicijnen of van de behandeling, en kosten van allerlei aard. De gezondheidstoestand kan specifiek worden benaderd door te focussen op één dimensie, of breed worden benaderd door verschillende dimensies tegelijk te onderzoeken. Bovendien kunnen verschillende ziektes verschillende aspecten van de gezondheidstoestand aantasten en individuele patiënten kunnen verschillende prioriteiten hebben ten aanzien van hun gezondheidstoestand. Daarom kunnen uitkomstmetingen geclassificeerd worden als specifiek of algemeen in elk van drie gebieden: gezondheidstoestanden, ziekten of aandoeningen, en patiënten (zie omslag van dit proefschrift). Een maat is specifiek in het gezondheidstoestand-gebied als het slechts één dimensie meet, en algemeen als het verschillende dimensies meet. Een maat is specifiek in het ziekte- of aandoening-gebied als het toepasbaar is op maar één aandoening (bijvoorbeeld spondylitis ankylopoetica). De maat wordt minder specifiek als het toepasbaar is op een groep van aandoeningen (bijvoorbeeld alle vormen van artritis, of alle vormen van kanker) en volledig algemeen als het toepasbaar is op alle mogelijke aandoeningen. Zo ook refereren maten die specifiek zijn in het patiënt-gebied naar individuele patiënten. Minder specifiek gerichte maten refereren naar subgroepen (bijvoorbeeld de ouderen) en algemene maten refereren naar alle mogelijk patiënten. Uitkomstmetingen kunnen algemeen zijn in één gebied en specifiek in een ander gebied.

Hoofdstuk 2 geeft een overzicht van uitkomstmetingen met betrekking tot de gezondheidstoestand die vandaag de dag worden gebruikt in spondylitis ankylopoetica. Van de 43 besproken studies werd in 79% gebruik gemaakt van maten die door de arts werden gemeten. In 67% werd gebruik gemaakt van maten die door de patiënt werden gerapporteerd. De meeste van de door de arts gemeten maten (67%) richtten zich op fysieke beperkingen (beweeglijkheid van de wervelkolom). De meeste van de door de patiënt gerapporteerde maten (65%) richtten zich op symptomen zoals pijn en stijfheid. De meest algemene maat, 'global

assessment' door arts of patiënt uitgedrukt in één item, werd in respectievelijk 16% en in 40% van de studies gerapporteerd. Andere algemene maten die de gezondheidstoestand meten werden maar af en toe gebruikt. Met name ontbrak het vaak aan de maten die door de patiënt worden gerapporteerd en die het fysieke functioneren van de patiënt meten. Er was slechts één studie die de prioriteiten van de patiënt ten aanzien van behandelingsrisico's rapporteerde en slechts drie studies omvatten een economische analyse. De rapportage van bijwerkingen van medicijnen of van de behandeling liet te wensen over. Dit literatuuroverzicht laat duidelijk zien dat sommige gebieden van de gezondheidstoestand niet worden gedekt met de gebruikte uitkomstmetingen en dat andere gebieden relatief veel aandacht krijgen.

De hoofdstukken 3 tot en met 8 behandelen prioriteiten van patiënten.
Hoofdstuk 3 geeft een inleiding tot utiliteitsmeting in de reumatologie. Het unieke van utiliteitsmaten is dat deze algemeen zijn in zowel het gezondheidstoestand-gebied als in het ziekte- of aandoening-gebied. In het patiënt-gebied echter is het specifiek wanneer utiliteiten van individuele patiënten worden gemeten, doch algemeen wanneer utiliteiten van de algemene bevolking worden verkregen. Met de utiliteit kunnen vergelijkingen worden gemaakt tussen verschillende aandoeningen en interventies. Dit hoofdstuk beschrijft twee methoden van utiliteitsmeting. Bij de eerste methode worden patiënten geclassificeerd in categorieën die gebaseerd zijn op antwoorden op vragen over het functioneren van patiënten (bijvoorbeeld, de 'Quality of Well-Being questionnaire' en de 'Sickness Impact Profile'). Bij de tweede methode van utiliteitsmeting wordt de patiënt zelf direct gevraagd een waardering te geven over zijn of haar algehele gezondheid. De vier meest gebruikte methoden (rating scale, standard gamble, time trade-off en willingness to pay) worden beschreven.

Hoofdstuk 4 beschrijft de bruikbaarheid van utiliteitsmeting met de rating scale en de standard gamble methode bij patiënten met spondylitis ankylopoetica (SA) of fibromyalgie (FMS). In het kader van twee gerandomiseerde gecontroleerde studies (trials) werd door getrainde interviewers de 'Maastricht Utility Measurement Questionnaire' (MUMQ) afgenomen bij 59 patiënten met SA ten tijde van de negen maanden followup en bij 86 patiënten met FMS op de baseline meting. De MUMQ is een Nederlandse aangepaste vertaling van de 'McMaster Utility Measurement Questionnaire' die door Bennett en Torrance werd ontwikkeld. De MUMQ omvat zowel de rating scale als de standard gamble methode. De bruikbaarheid van beide methodes bij zowel patiënten met SA als patiënten met FMS was over het algemeen bevredigend. Alle patiënten voltooiden het interview. Vier (2.8%) van alle 143 patiënten gaven inconsistente antwoorden: één op de rating scale en drie op de standard gamble. De duur van het interview was ongeveer 9-12 minuten (SD 3.6-5.4) met de rating scale en ongeveer 12-14 minuten (SD 2.9-3.8) met de standard gamble methode. De vier-week test-hertest betrouwbaarheid voor de gezondheidstoestand van de patiënt, uitgedrukt als intraclass correlatie coëfficiënt, was 0.56 voor de rating scale en 0.66 voor de standard gamble. Op de rating scale (0-100) waardeerden patiënten met SA hun eigen gezondheidstoestand aanzienlijk hoger dan patiënten met FMS (SA:69 en FMS:54). Echter, de standard gamble utiliteitswaarden (0-1) waren voor beide aandoeningen ongeveer gelijk doch op een hoger niveau (SA: 0.86 en FMS:0.83). Deze data ondersteunen de opvatting dat het resultaat van de utiliteitsmeting afhankelijk is van de methode die gekozen wordt om het welbevinden van de patiënt te meten.

In **hoofdstuk 5 en 6** worden utiliteiten van patiënten met SA of FMS die verkregen zijn met de rating scale methode, vergeleken met utiliteiten van deze patiënten die verkregen zijn met de

standard gamble methode. Vervolgens worden deze utiliteiten gerelateerd aan andere uitkomstmaten, en wordt de gevoeligheid voor verandering van utiliteiten vergeleken met de gevoeligheid voor verandering van andere uitkomstmaten. De MUMQ werd twee keer afgenomen bij beide patiëntengroepen: ten eerste op baseline en ten tweede na zes (FMS) of negen (SA) maanden followup. Bij SA werd de analyse beperkt tot slechts 59 van de 135 patiënten die door dezelfde interviewer bij beide metingen werden gezien. Bij FMS werden alle 73 patiënten door dezelfde interviewer gemeten op beide meetmomenten. De construct validiteit leek bij zowel patiënten met SA als patiënten met FMS hoger voor de rating scale dan voor de standard gamble methode. Voor de rating scale methode varieerden de Spearman correlatie coëfficiënten van 0.25 tot 0.68 bij patiënten met SA en van 0.23 tot 0.62 bij patiënten met FMS; voor de standard gamble methode varieerden deze van 0.23 tot 0.36 bij beide aandoeningen. Regressie analyses lieten zien dat de 'global assessment' van de patiënt 41 - 59% van de rating scale utiliteiten verklaarden en 10 - 11% van de standard gamble utiliteiten bij patiënten met SA of FMS. Deze bevindingen ondersteunen de gedachte dat utiliteiten verkregen via de rating scale methode de 'global assessment' dichter benaderen. Het methodologische voordeel van gestandaardiseerde rating scale utiliteitsmeting boven niet-gestandaardiseerde 'global assessment' is dat utiliteiten numerieke waarderingen opleveren die het mogelijk maakt patiënt-uitkomsten van verschillende aandoeningen of patiënt-uitkomsten die resulteren van verschillende interventies te vergelijken. Bij beide aandoeningen correleerden veranderingen in rating scale utiliteiten beter met veranderingen in andere uitkomstmaten dan veranderingen in standard gamble utiliteiten.

Hoofdstuk 7 behandelt diverse methodologische aspecten van utiliteitsmetingen met de rating scale en de standard gamble methode. Deze methodologische aspecten werden gedestilleerd uit twee gerandomiseerde klinische trials waaraan respectievelijk 85 FMS patiënten en 144 SA patiënten deelnamen. In beide trials vond één baseline meting en twee followup metingen plaats waarin patiënten gevraagd werden hun waarderingen uit te spreken over hun eigen gezondheidstoestand en enkele hypothetische gezondheidstoestanden (licht, matig, en ernstig ziek).
De standard gamble methode bleek consistent tot hogere scores te leiden dan de rating scale, zowel voor de eigen gezondheidstoestand als voor de hypothetische gezondheidstoestanden. De drie-maand test-hertest betrouwbaarheid voor de hypothetische toestanden uitgedrukt als intraclass correlatie coëfficiënt varieerde van 0.24 tot 0.33 voor de rating scale en van 0.43 tot 0.70 voor de standard gamble. Hoewel de reproduceerbaarheid niet hoog was, bleven de groepsgemiddelden in de loop van de tijd relatief constant. De gemiddelde standard gamble score lijkt afhankelijk te zijn van de wijze waarop de metingen worden uitgevoerd. Utiliteiten gemeten met de zogenaamde 'chained gambles' waren significant hoger dan utiliteiten gemeten met de 'basic reference gamble'. Op het individuele niveau werden inconsistente antwoorden gevonden. Meer dan 70% van deze inconsistente antwoorden viel echter binnen de grenzen van de meetfout, die varieerde van 0.11 tot 0.13 op de standard gamble (0-1 schaal) en van 8 tot 10 op de rating scale (0-100 schaal). Het grote aantal negatieve utiliteiten voor de hypothetische toestand van een ernstig zieke - de toestand die als ankerpunt in de 'chained gambles' was gebruikt - en de waarde van de negatieve utiliteiten (tot -19) onderstrepen de noodzaak betere methoden te ontwikkelen voor het meten van negatieve utiliteiten en het hanteren van deze utiliteiten in 'Quality Adjusted Life Years' (QALY) berekeningen.

Hoofdstuk 8 beschrijft een andere manier om prioriteiten van patiënten vast te stellen en wel met behulp van de 'Problem Elicitation Technique (PET) questionnaire'. Deze vragenlijst behandelt alleen die activiteiten die direct door de aandoening worden beperkt en als belangrijk worden ervaren door de patiënt. De PET vragenlijst werd door getrainde interviewers afgenomen bij 134 patiënten met SA en 73 patiënten met FMS op de baseline meting en na respectievelijk negen en zes maanden followup. Op iedere meting werd een PET score berekend.
Patiënten met FMS gaven meer problemen aan (gemiddeld 6.8) dan patiënten met SA (gemiddeld 4.4). Bovendien konden patiënten met SA vaker geen probleem noemen op de baseline meting dan patiënten met FMS (10% in vergelijking tot 1%). De PET score verbeterde van 14.9 naar 11.3 ($p=0.0001$) bij patiënten met SA, maar veranderde niet (van 21.8 naar 21.1) ($p=0.24$) bij patiënten met FMS.
De construct validiteit van de PET score werd getoetst middels correlaties tussen de PET score en andere effectmaten. De PET score correleerde statistisch significant met de 'Arthritis Impact Measurement Scale' (AIMS), utiliteiten, 'Sickness Impact Profile' (SIP), 'Health Assessment Questionnaire' (HAQ), pijn, stijfheid, en de algehele gezondheidstoestand van de patiënt bij zowel patiënten met SA als FMS (r's varieerden van 0.22 tot 0.66). In een multiple regressie analyse verklaarde pijn 29% van de variantie in de PET scores bij patiënten met SA. Bij FMS verklaarde de 'global assessment' van de patiënt 39% van de totale variantie in PET scores, terwijl pijn nog eens 15% verklaarde. Veranderingen in PET scores correleerden significant ($p<0.05$) met veranderingen in AIMS, utiliteiten, pijn, stijfheid en de algehele gezondheidstoestand van de patiënt. Dit betrof zowel patiënten met SA als patiënten met FMS (r's varieerden van 0.22 tot 0.51). Veranderingen in PET scores werden voor 6% verklaard door veranderingen in pijn bij patiënten met SA en voor 35% door veranderingen in pijn en subjectieve gezondheid bij patiënten met FMS. Het meten van de gevoeligheid voor verandering toonde aan dat de efficiëntie van de PET score 0.6 was bij patiënten met SA en 0.09 bij patiënten met FMS. Vergeleken met andere uitkomstmaten is dit redelijk voor patiënten met SA, maar laag voor patiënten met FMS.
Over het algemeen was het goed mogelijk om prioriteiten van de patiënt in beeld te krijgen. Bij zowel patiënten met SA als FMS was de construct validiteit van de PET vragenlijst bevredigend. De PET was veel gevoeliger voor verandering bij patiënten met SA dan bij patiënten met FMS.

Hoofdstuk 9 bespreekt de kosten-effectiviteit van groepsoefentherapie bij patiënten met SA. Vergeleken met thuis oefenen levert groepsoefentherapie na negen maanden een extra verbetering op van 7% in beweeglijkheid van de wervelkolom, 5% extra verbetering in fitheid en 28% extra verbetering in globale gezondheid voor een bedrag van 1001 gulden per jaar. De directe medische kosten (bezoeken aan de huisarts, specialist en paramedicus, het gebruik van medicijnen en het aantal dagen opgenomen in het ziekenhuis) worden met 230 gulden per jaar gereduceerd. De effecten van groepsoefentherapie kosten dus 771 gulden per jaar, oftewel bijna 20 gulden per week.

Hoofdstuk 10 plaatst enkele kanttekeningen bij het onderzoek beschreven in dit proefschrift.

DANKWOORD

Velen hebben bijgedragen aan de totstandkoming van dit proefschrift. Allereerst wil ik alle patiënten bedanken voor hun medewerking aan de onderzoeken. Marco Aarts, Rob de Bie, Cootje Braakman, Carla Jonker, en Annemiek Mackaay toonden hun inzet bij het afnemen van de interviews bij de patiënten. Hubert Schouten was behulpzaam met zijn deskundige statistische adviezen. Paulien Copper corrigeerde mijn Engelstalige teksten en Marieke Eelkman Rooda mijn Nederlandstalige teksten.

De (oud)-secretaresses van onze werkgroep reumatologie Peggy Renckens-Lamkin, Yolanda Soons, Lilian Stassen, en Mieke Hamers zorgden voor een prettige werksfeer en de nodige ondersteuning.

Tiny Wouters ben ik zeer erkentelijk voor de enthousiaste en deskundige wijze waarop zij de lay-out van dit proefschrift heeft verzorgd.

I gratefully acknowledge the support of Kathryn Bennett for providing us the McMaster Utility Measurement Questionnaire. Thank you for the phone calls and faxes in which we discussed utility measurement results. Thank you for the very pleasant co-operation.
I also acknowledge the support of Claire Bombardier and Rachelle Buchbinder for providing us the Problem Elicitation Technique questionniare.

Maarten Boers wil ik bedanken voor zijn hulp bij de tot standkoming van hoofdstuk 2 en voor de vele boeiende discussies op het gebied van utiliteitsmeting.

De studie 'groepsoefentherapie bij patiënten met spondylitis ankylopoetica' werd uitgevoerd door Alita Hidding. De studie 'fitness en biofeedback training bij patiënten met fibromyalgie' werd uitgevoerd door Paulien Bolwijn en Marijke van Santen-Hoeufft. Ik kijk met veel plezier terug naar de goede samenwerking en dat jullie niet vergaten dat de utiliteitsinterviews, PET- en kosten-vragenlijsten ook afgenomen moesten worden. Paulien, bedankt voor de leuke tijd die wij als SAS-girls hebben doorgebracht en niet te vergeten voor de vele gezellige uren waarin we ervaringen uitwisselden over onze taken buiten het werk.

Ineke Blaauw en Zuzana de Jong-Straková wil ik bedanken voor alle steun en betrokkenheid op

en buiten het werk. Ineke, jij hebt me wegwijs gemaakt in Maastricht, bij buitenlandse congressen en in hoe te promoveren. Zuzana, samen konden wij bijna alles relativeren met een lach.

Prof Dr EKA van Doorslaer wil ik bedanken voor zijn enthousiaste begeleiding op het gebied van de medical technology assessment. Beste Eddy, je wist altijd tijd vrij te maken om mij te helpen met de vele vragen die ik had over het meten van utiliteiten en kosten. We hebben ons verbaasd over de vaak verrassende antwoorden van patiënten.
Maureen Rutten-van Mölken wil ik bedanken voor haar bijdragen aan het ontwikkelen van de utiliteitsmetingen. Als pioniers op het gebied van de utiliteitsmetingen bij patiënten hebben we samen met Eddy en Sjef vele discussies gevoerd over de methoden, resultaten en de zin en onzin van deze metingen. Van jou en van deze discussies heb ik veel steun ondervonden.

Prof Dr Sj van der Linden, beste Sjef: je was de inspirator van dit proefschrift. De begeleiding die je me gaf was stimulerend en motiverend. Vanaf het begin bood je me alle mogelijkheden om deskundig te worden op het gebied van het meten van gezondheids-gerelateerde kwaliteit van leven in de reumatologie. Cursussen, congressen en bezoeken aan Hamilton behoorden tot die mogelijkheden. Bedankt voor alle steun op dit moeilijke doch interessante werkterrein.

Anita en Magda van kindercreche 'het Grummelke', Windy, Colinde, Ilse, Saskia, Nathalie, Rinie en Marga van kinderdagverblijf 'de Hummelhof', Nanny en Sheila bedank ik voor de goede zorgen voor Floris en Ruben.

Mijn ouders bedank ik voor alle mogelijkheden die ze me geboden hebben.

Lieve Berno, Floris en Ruben, jullie zijn uniek en mijn grootste stimulans.

CURRICULUM VITAE

Carla Bakker werd geboren op 16 februari 1962 te Geleen. In 1980 behaalde zij het eindexamen atheneum-ß aan de Albert Schweitzer Scholengemeenschap in Geleen. In dat jaar startte zij met de opleiding voor Fysiotherapie te Leiden, die in 1984 met een diploma werd afgesloten. Na haar afstuderen was zij gedurende een jaar werkzaam als vervangend fysiotherapeut in een praktijk voor fysiotherapie te Leiden. Tevens startte zij in 1984 met de studie Gezondheidswetenschappen aan de Rijksuniversiteit Leiden, waar zij in 1988 afstudeerde in de Klinische Epidemiologie bij Prof.Dr. JP Vandenbroucke. Vervolgens was zij van 1989 tot 1994 werkzaam als toegevoegd wetenschappelijk onderzoeker bij de vakgroep Interne Geneeskunde van het Academisch Ziekenhuis Maastricht binnen de werkgroep Reumatologie onder leiding van Prof.Dr. JMJP van der Linden. Zij is gehuwd met Berno Gerts en heeft twee zonen, Floris en Ruben.